1

Two Dimensional Linkage Study Methods and Related Inventions

2 The present patent application is a continuation-in-part of U.S. Patent Application 09/947,768 (filed 5

3 SEPT 2001). And 09/947,768 claims priority from US Provisional 60/230570 (filed 9/5/2000). Patent

4 Application 09/947,768 is a continuation-in-part of U.S. Patent Application 09/623,068 (filed 26 AUG

5 2000). The present patent application is also a continuation-in part of Patent Application 09/623,068

6 (filed 26 AUG 2000). Application 09/623,068 claims priority from PCT/US99/04376 filed (2/26/99).

7 PCT/US99/04376 claims priority from US Provisional applications: 60/076182 filed 27Feb1998,

8 60/086947 filed 27May1998, 60/076102 filed 26 Feb 1998 and 60107673 filed 7 Nov 1998 and

9 60/326,331 filed 1 Oct 2001. Each of the following patent applications are incorporated herein by

10 reference in their individual entireties: U.S. Provisional Patent Application 60/230570 and 60/326,331,

11 PCT/US99/04376, U.S. Patent Application 09/623,068, and U.S. Patent Application 09/947,768.

12 The reader's attention is directed to the following documents or papers each of which is open to the

13 public and each of which is incorporated by reference herein in their entirety: (1) McGinnis, Ewens &

14 Spielman, Genetic Epidemiology 1995 ; 12(6) : 637-40. (2) RE McGinnis Annals of Human Genetics vol

15 62, pp. 159-179

Technical Field

17 Versions of the present invention are in the field of molecular biology, some versions are specifically in

18 the area of finding the chromosomal location of genes that cause genetic characteristics such as

19 human disease.

20

Background

22

Introduction

24 Conventional linkage study techniques have limited power to localize trait causing genes (trait causing

25 polymorphisms) of modest effect, such as many human disease polymorphisms. The two-dimensional

26 linkage study techniques of this application are powerful new techniques for localizing genes

27 (polymorphisms) especially of modest effect.

28 Chromosomes, heredity, genes, markers and alleles

29 Chromosomes are large molecules that carry the information for the inheritance of physical (genetic)

30 characteristics or traits. In human beings for example, parents pass a copy of half of their chromosomes

31 to their offspring during reproduction. By doing this, each parent passes some of his or her physical

32 characteristics to his or her offspring. Any chromosome of a living creature is made of a large string-like

33 molecule of DNA. Chromosomes are essentially very long strings of DNA. Genes are small pieces of a

34 chromosome that cause or determine inherited genetic characteristics. (In this application, the term

35 gene means a polymorphism that determines a genetic characteristic; the term does not mean an entire

36 gene structure with a promoter region, introns, etc..) Markers are any segment of DNA on a

37 chromosome which can be identified and whose chromosomal location is known (at least to some

38 extent). Markers are like milestones along the very long string-like molecule of DNA which makes up a

39 chromosome. Both a gene and a marker can come in different forms on different chromosomes. These

1 different forms are known as different alleles and when a gene or marker comes in different forms it is

2 said to be "polymorphic". For example, a bi-allelic marker comes in two (bi) different forms.

3 Linkage

4 If a gene allele and a marker allele occur as part of the genetic makeup of individuals more frequently

5 than would be expected on the basis of chance, then it is possible to infer that the gene and the marker

6 are linked. If a gene allele and a marker allele are inherited together more frequently than would be

7 expected if the gene and the marker were on different chromosomes, then it is possible to infer that the

8 gene and the marker are linked. Linkage of a gene and a marker usually occurs because the gene and

9 the marker are close together on a chromosome. There are different degrees of linkage. Establishing

10 linkage, especially strong linkage, between a gene and a marker can be very valuable. This is

11 especially true if the precise location and other characteristics of the gene are not known. By

12 establishing linkage, especially strong linkage, between a known marker and an unknown gene it is

13 possible to locate the gene near to the chromosomal location of the known marker. This can be very

14 valuable if the gene is an important gene, such as a disease causing gene, and can help cure the

15 disease.

16 Linkage Studies

17 Linkage studies are a method of establishing linkage between a marker and a gene or genes. Linkage

18 studies are used to statistically correlate the occurrence of a genetic characteristic such as a disease

19 (caused by a gene or genes) with a marker on a chromosome. One way this is done is by statistically

20 correlating a specific allele of a marker with a genetic characteristic for a set of individuals by showing

21 that individuals with the characteristic inherit the marker allele more often than individuals without the

22 characteristic. The set of individuals is usually referred to as a sample of individuals. An example of a

23 sample of individuals are people with a disease and similar people (matched controls) without the

24 disease. Another example of a sample of individuals is a group of people, some of whom have the

25 same disease; each of the people in the group being related to one or more of the other people in the

26 group (i.e. families, sibships, pedigrees). The presence or absence of a marker allele in the

27 chromosomal DNA of each individual is usually determined by genotype data at the marker for each

28 individual.

29 There are different types of linkage study techniques, using different types of samples and different

30 statistical measures of the correlation of a marker and a genetic characteristic. One example of a type

31 of linkage study technique is the affected sib pair (ASP) test. Another example is the transmission

32 disequilibrium test (TDT), which is an association based linkage test. This is a dynamic, changing area

33 within the field of human genetics.

34 Linkage Studies and the "Scanning" of Chromosomal Regions

35 There are significant advantages in using several markers simultaneously to perform a linkage study

36 with a genetic characteristic and a sample of individuals, especially when the relative positions of the

37 markers on a chromosome are known. Such a linkage study allows searching for statistical evidence of

38 linkage between markers in one or more regions of a chromosome or chromosomes and the gene or

39 genes that determine the genetic characteristic. The results of the study for each marker can then be

1   compared with the results for other markers, knowing the relative chromosomal positions of all the

2   markers in the study. In this way, regions of a chromosome or even whole chromosomes can be

3   "scanned" for evidence of linkage to a gene or genes causing a genetic characteristic. The relative

4   positions of markers on chromosomes of a species of creatures is given by various kinds of

5   chromosomal maps for the species. (There are several different kinds of marker maps, i.e. physical

6   maps, genetic maps, radiation hybrid maps, etc.)

7   <u>Sets of Markers for Linkage Studies and "Scanning" Chromosomes</u>

8   An appropriate set of markers from a region of a chromosome can be chosen so that the region can be

9   "scanned" for evidence of linkage of markers in the region to a gene or genes that cause a genetic

10   characteristic. As explained above, this scanning is done by using the markers in linkage studies.

11   Strong positive evidence for linkage of the markers (from the scanned chromosomal region) to a gene

12   or genes responsible for a characteristic or trait is strong evidence that a trait-causing gene or genes is

13   located within the chromosomal region.

14   <u>Conventional Techniques for Choosing Sets of Markers to Scan Chromosomes with Linkage Studies</u>

15   Conventional techniques choose sets of markers to scan a chromosomal region by choosing markers

16   according to each marker's chromosomal location within the region. In a set of microsatellite markers

17   described in 1994 for use in linkage studies, the markers were approximately evenly spaced, with

18   average spacing between markers being 13 centiMorgans. The markers were distributed approximately

19   evenly across the entire human genome (all human chromosomes) and were also selected because

20   genotype data at the markers for individuals could be obtained by a semi-automated method.[1] A recent

21   (1998) linkage study of the disease schizophrenia used a set of 310 microsatellite markers distributed

22   approximately evenly across the entire human genome with average spacing of 11 centiMorgans

23   between markers.[2] In a recent (1998) simulation of linkage studies to defend the practice of two-stage

24   genome scanning, markers were spaced evenly every 10 cM(centimorgans) in an initial, sparser, first

25   stage scan and evenly every 1 cM in a followup, denser, second stage scan.[3] Following up positive

26   linkage study results from chromosomal regions in a sparse, first stage scan with a second, denser

27   scan that focuses on studying the regions with positive first-stage results is a common technique. In

28   these conventional studies, as is common, markers were chosen to be about evenly spaced across the

29   chromosomal regions studied. In this manner, as is conventional, a one dimensional structure such as

30   an entire genome, a chromosome or a region of a chromosome is "covered" by markers in order to

31   scan the entire genome, chromosome or chromosomal region with a linkage study. (These conventional

32   techniques[1,2,3] are not admitted to be prior art by their mention in this background.) (There is a

33   possibly confusing, double meaning, of the term "marker map". It should be noted that a set of markers

34   distributed along a chromosomal region, chromosome, or genome for linkage studies is also sometimes

---

[1] Reed, et.al.: Chromosome-specific microsatellite sets for flourescence-based, semi-automated genome mapping. Nature Genetics, July 1994; vol. 7: pp. 390-395.

[2] Levinson, et.al.: Genome Scan of Schizophrenia. Am J Psychiatry, June 1998; vol. 155: pp. 741-750.

[3] Kruglyak, et. al.: Linkage Thresholds for Two-stage Genome Scans. Am J Hum Genet, 1998, vol. 62: pp. 994-996.

1   referred to as a "marker map" for use in chromosomal scanning by linkage studies. In addition,

2   chromosomal or genetic maps of markers are also referred to as "marker maps".)

3   **Conventional Techniques for Choosing Sets of Markers to Scan Chromosomal Regions are**

4   **Essentially One Dimensional**

5   Because DNA is a stringlike molecule, a chromosomal region(s), chromosome(s) and genome are

6   essentially one dimensional in terms of the chromosomal location of markers and genes. As has been

7   stated, conventional linkage study techniques scan a chromosomal region(s), chromosome(s) or

8   genome by using markers distributed approximately evenly along the length of the chromosomal

9   region(s), chromosome(s) or genome respectively. These conventional techniques focus primarily on

10   the chromosomal location of markers used in a scan. These conventional techniques have an

11   essentially one dimensional perspective.

12   Population Frequency of Marker Alleles and Gene Alleles

13   As described, chromosomal location of each marker is an important and unique characteristic of each

14   marker and marker allele. Another characteristic of each polymorphic marker and each of the marker's

15   alleles is the population frequency of each marker allele. A population is a group (usually a large group)

16   of individuals. A population frequency of a particular marker allele is the proportion of individual

17   chromosomes in a population in which the particular marker occurs as the particular marker allele. For

18   any bi-allelic marker, knowing the least common allele frequency of the marker establishes both of the

19   allele frequencies of the marker. This is because the two allele frequencies of a bi-allelic marker sum to

20   1. Each gene allele also has a population allele frequency or allele frequency for short. Thus, each

21   gene allele has a particular chromosomal location and allele frequency (for a particular population). In

22   the case of an unknown gene, the gene's chromosomal location and allele frequencies are not

23   specifically known.

24   Marker Allele Population Frequency in Conventional Linkage Study Scans

25   It is important to note that little attention was paid to the population allele frequencies of the markers

26   used in the conventional linkage scans cited above. In the two studies cited above under conventional

27   scanning techniques[1, 2], marker allele frequency is referred to only peripherally as average marker

28   heterozygosity, which is related to average marker allele frequency and the number of alleles (2, 3, 4, 5,

29   etc.) at each marker. In the simulated scan cited above[3], *the markers are stipulated to have four alleles*

30   *that all have exactly the same allele frequency of 0.25* (heterozygosity 0.75). *It is important to note*

31   *that while the chromosomal location of the markers in all these conventional scans was*

32   *systematically varied over the entire genome (all the human chromosomes), nothing was said*

33   *about systematically varying the allele frequencies of the markers in any of the scans.* This is

34   typical of conventional linkage study scans of genomes, chromosomes and chromosomal regions.

35   A Conventional View Of Bi-allelic Markers And Linkage Studies

36   We cite here a well known reference that discusses the conventional view of bi-allelic marker

37   usefulness in linkage scans of chromosomes. In 1997 Kruglyak carried out computer simulations of the

1   "information content" of markers that are part of various different marker maps.[4] For bi-allelic markers

2   his results showed that the optimum allele frequencies for bi-allelic markers used in linkage studies is

3   0.5/0.5 in order to achieve the greatest information content. However, allele frequency patterns other

4   than the optimum 0.5/0.5 for bi-allelic markers gave acceptable levels of information content depending

5   on the density of the marker map (or set of markers) chosen for the linkage study.

6   There are some important observations regarding this reference.[4] First, **there is no advantage noted**

7   **in this reference for choosing bi-allelic markers so that the set of chosen markers (or marker**

8   **map) used for linkage studies is such that the markers systematically vary in allele frequency.**

9   Thus, just as in the recent conventional linkage study scans cited above, there is no definite thought to

10   using markers of systematically varying allele frequencies. The greatest information content is given by

11   bi-allelic markers with allele frequencies close to the optimum of 0.5/0.5. Given the density of

12   reasonably polymorphic SNPs predicted in this reference, at least one every 1 kb or 1,000 per cM, it is

13   probable that even for quite dense maps, there will be so many acceptable SNPs available, that all of

14   the SNPs in an appropriate marker map could have the optimum allele frequencies of approximately

15   0.5/0.5. Secondly, bi-allelic markers with lower least common allele frequencies, less than 0.3(0.7/0.3)

16   or 0.2(0.8/0.2), are viewed unfavorably for linkage studies in this reference. Thirdly, the early version of

17   the criterion of "information content" of markers used in this reference was based on sib pair analysis

18   and the later, current version of the criterion, does not depend on any particular test for linkage.[5, 6]

19   **Thus, the criterion of information content in this reference, has never specifically employed the**

20   **TDT (transmission disequilibrium test) or any association based test, whereas the two-**

21   **dimensional linkage study techniques of this application are based on a completely different**

22   **perspective of using association based tests.** (This reference[4] is not admitted to be prior art with

23   respect to the present invention by it's mention in this background.)

24   <u>Increased Power of the TDT (transmission disequilibrium test)</u>

25   Characteristics of a new type of linkage test, the TDT (transmission disequilibrium test), were described

26   in 1993. The inventor, R.E.McGinnis, was one of the authors of this reference.[7] In 1996, Risch and

27   Merikangas argued that conventional linkage analysis has limited power to detect genes of modest

28   effect. And Risch and Merikangas attempted to illustrate the increased power of association based

29   linkage tests such as the TDT over other types of conventional linkage tests.[8] However, Risch and

30   Merikangas' analysis was criticized by Muller-Myhsok and Abel as being based on the optimal

31   assumption that the analyzed allele was the disease allele itself. Muller-Myhsok and Abel concluded

---

[4] Kruglyak: The use of a genetic map of biallelic markers in linkage studies. Nature Genetics, September 1997, vol.17, pp. 21-24.

[5] Kruglyak, et. al.: Complete Multipoint Sib-Pair Analysis of Qualitative and Quantitative Traits. Am J Hum Genet, 1995, vol. 57: pp. 439-454.

[6] Kruglyak, et. al.: Parametric and Nonparametric Linkage Analysis: A Unified Multipoint Approach. Am J Hum Genet , 1996, vol. 58, pp. 1347- 1363.

[7] Spielman, R.S., McGinnis, R.E., Ewens, W.J.: Transmission Test for Linkage Disequilibrium: The Insulin Gene Region and Insulin-dependent Diabetes Mellitus(IDDM). Am J Hum Genet, 1993, vol. 52, pp. 506-516.

1   that researchers should be aware that the power of association studies such as the TDT can be greatly

2   diminished in more common, less optimal situations.[9] In their response to Muller-Myshok and Abels'

3   letter, Risch and Merikangas essentially agreed with the logic of Muller-Myshok and Abels' criticism.

4   Risch and Merikangas stated that to a large extent, the expectation with respect to linkage

5   disequilibrium across the genome is uncharted territory.[10] (None of the references in this paragraph [7,8,]

6   [9,10] is admitted to being prior art with respect to the present invention by their mention in this

7   background.)

8   <u>More Detailed Studies of the Power of the TDT</u>

9   The inventor, R.E.McGinnis, has done extensive investigations on the power of the TDT. His

10   observations and calculations of the increased power of the TDT in many situations have been

11   published.[11] In this paper a general framework for determining the power of the TDT in many different

12   situations is presented. The analysis of Risch and Merikangas [8] and others is shown by the inventor to

13   be a special case of his general framework. His observations and calculations published in this paper

14   have shown that the TDT has increased power in more common, less optimal situations as well as the

15   less common, optimal situation cited by Muller-Myshok and Abel [9]. As opposed to the observation of

16   Muller-Myhsok and Abel, the inventor's calculations indicate that association tests such as the TDT

17   have increased power in typical situations even when the ratio m/p departs significantly from unity and,

18   or the linkage disequilibrium between the analyzed (marker) allele and disease polymorphism is only

19   half its maximum possible value. The inventor arrived at these conclusions independently and did not

20   derive them from others.

21   ***<u>A Major Conclusion Drawn by the Inventor about the TDT and Linkage Studies: Using Bi-allelic</u>***

22   ***<u>Markers of Systematically Varying Allele Frequencies Increases the Power of Linkage Studies</u>***

23   ***<u>Using the TDT</u>***

24   The inventor's calculations and observations about the increased power of the TDT in more common,

25   less optimal situations led him to the conclusion that the power of linkage studies using the TDT is

26   greatly increased under some conditions. Under some conditions, the power of the TDT in a linkage

27   study using bi-allelic markers is greatly increased when each of one or more of the bi-allelic markers

28   used in the study fulfill two criteria: (1) the allele frequencies of each of the one or more of the bi-allelic

29   markers are similar (but not necessarily the same, or even approximately the same) as the allele

30   frequencies of an unknown bi-allelic gene causing a disease under study; and (2) each of the one or

31   more bi-allelic markers is in some degree of linkage disequilibrium with the gene. Thus for a typical

32   linkage study using bi-allelic markers and the TDT, *to increase the likelihood of conditions occurring*

33   *that increase the power of the TDT in the linkage study, the bi-allelic markers used in the study are*

34   *chosen so that the least common allele frequencies of the markers vary systematically over a range or*

[8] Risch, N. and Merikangas, K.: The Future of Genetic Studies of Complex Human Diseases. Science, 13 September 1996, vol. 273, pp. 1516-1517.

[9] Muller-Myshok, B. and Abel, L.: Technical Comments: The Future of Complex Diseases. Science, 28 February 1997, vol. 275, pp. 1328-1329.

1   *subrange of least common allele frequency.* This major conclusion of the inventor's research is quoted

2   directly from his unpublished manuscript that was included with previously filed U.S. Provisional Patent

3   Applications:"This example is typical and highlights perhaps the most important finding of this paper;

4   namely the importance of using bi-allelic markers with heterozygosity similar to that of a bi-allelic

5   disease gene. Indeed, since a majority of susceptibility loci may be bi-allelic, the judicious use of bi-

6   allelic markers of both high, medium and low heterozygosity may be crucial in order to detect and

7   replicate linkages to loci conferring modest disease risk." (page 25) (In this context the phrase "bi-allelic

8   markers with heterozygosity similar to that of a bi-allelic disease gene" is essentially equivalent to "bi-

9   allelic markers with individual allele frequencies similar to those of a bi-allelic disease gene" and "bi-

10   allelic markers of both high, medium and low heterozygosity " is essentially equivalent to the phrase "bi-

11   allelic markers whose least common individual allele frequencies are high, medium and low".)

12   **Systematically Varying Both Marker Chromosomal Location and Marker Allele Frequency of Markers in**

13   **Linkage Studies**

14   The inventor's calculations and observations have demonstrated the increased power of the TDT in

15   more common, less optimal situations when a bi-allelic marker and bi-allelic gene have (1) similar but

16   not identical allele frequencies and (2) the marker and gene are in some degree of linkage

17   disequilibrium. Thus, for a typical linkage study using bi- allelic markers and the TDT, *to increase the*

18   *likelihood of both criteria (1) and (2) occurring for one or more markers, so as to increase the power of*

19   *the TDT in the linkage study, the bi-allelic markers used in the study are chosen so that the least*

20   *common allele frequencies of the markers vary systematically over a range or subrange of least*

21   *common allele frequency AND the chromosomal location of the markers vary systematically over one or*

22   *more chromosomes or chromosomal regions. And the bi-allelic markers are chosen so that the*

23   *markers' chromosomal locations and least common allele frequencies vary systematically in an*

24   *essentially independent manner.*

25   **Two-dimensional Linkage Study Techniques**

26   As has been stated, conventional linkage study scanning techniques use markers that are distributed

27   approximately evenly in the dimension of chromosomal location. These conventional, one dimensional,

28   scanning techniques focus primarily on the chromosomal location of markers used in a scan and give

29   little attention to the dimension of allele frequency.[1, 2, 3]

30   One of the main implications of the inventor's work is to use a set of bi-allelic markers for a typical

31   linkage study using the TDT (or other association-based linkage test) wherein the chromosomal

32   locations and least common allele frequencies of the markers in the set systematically vary in an

33   essentially independent manner over the dimensions of chromosomal location and least common allele

34   frequency respectively. This is equivalent to using a set of bi-allelic markers for a linkage study scan

35   wherein the set of markers systematically scan or "cover" a two-dimensional region having dimensions

36   of chromosomal location and least common allele frequency. (Such a two-dimensional region can be

37   thought of as an area in an x-y plot or a group of squares on a chessboard.)

---

[11] McGinnis, R.E.: Hidden Linkage: Comparison of the affected sib pair (ASP) test and transmission disequilibrium test (TDT). Annals of Human Genetics, 1998, vol. 62, pp. 159-179.

1   In addition, the inventor's calculations and observations indicate that bi-allelic markers having least

2   common allele frequencies less than 0.3, 0.2 or even less than 0.1 have an important place in linkage

3   studies using association based linkage tests. This is markedly different than Kruglyak's information

4   content evaluation of bi-allelic markers for use in linkage studies, in which bi-allelic markers with least

5   common allele frequencies less than 0.3 or 0.2 are viewed unfavorably.[4]

6   *In addition, the two-dimensional linkage study techniques do not necessarily favor using markers in a*

7   *scan that are about evenly spaced along a chromosome as in the conventional techniques. **This is***

8   ***because conventional techniques suffer from a kind of one dimensional view or lack of depth***

9   ***perception.*** **In the conventional techniques, a marker can look very close to a gene's location in**

10   **terms of chromosomal location, but the marker can be very far from the gene's location in the**

11   **new two dimensional view used by versions of the invention.**

12   **It is as if the conventional 1D techniques look at a chessboard from on edge. Markers and a**

13   **gene which are on different squares of the board, but in the same column of squares, look very**

14   **close to each other when the board is looked at from on edge. But when the board is looked at**

15   **from the top in 2D, two dimensions, markers which looked very close to each other and the**

16   **gene before (when looking from on edge) can be seen to be very far from the gene.**

17   <u>Further Implications of the Two-dimensional Linkage Study Perspective</u>

18   These two-dimensional techniques work when multiple genes cause a genetic characteristic and are

19   effective in searching for these genes. A two-dimensional bi-allelic marker "covering" or scanning

20   approach also increases the power of linkage studies using other association based linkage tests such

21   as the AFBACmethod, the haplotype relative risk (HRR) method [12], and comparison of marker allele

22   frequencies in disease cases and unrelated controls[13]. These references[12, 13] are not admitted to being

23   prior art with respect to the present invention by their mention in this background.)

24   <u>Patents That May Be Helpful In Starting A Search Of The Background</u>

25   Some patents that are in the same general areas as versions of the invention are cited here: US Patent

26   Number 5,667,976 Solid supports for nucleic acid hybridization assays. Published International

27   Application WO 98/20165 Biallelic Markers. Published International Application WO 98/07887 Methods

28   for treating bipolar mood disorder associated with markers on chromosome 18 p. US Patent Number

29   5,552,270 Methods of DNA sequencing by hybridization based on optimizing concentration of matrix-

30   bound oligonucleotide and device for carrying out same. No patent in this paragraph is admitted to

31   being prior art with respect to the present invention by it's mention in this background.

32   <u>**Summary**</u>

33   **Versions of the invention use bi-allelic markers that "cover" or are distributed approximately**

34   **evenly (or systematically) over two-dimensional regions.** These regions have the two dimensions

35   of chromosomal location and least common allele frequency.

---

[12] Falk CT and Rubenstein P: Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Annals of Human Genetics, 1987, vol. 51, pp. 227-233.
[13] Bell GI, Horita S and Karam JH: A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. Diabetes, 1984, vol 33, pp. 176-183.

1 **Conventional techniques suffer from a kind of one-dimensional lack of depth perception.** (They

2 also favor bi-allelic markers with least common allele frequencies near 0.5.) Two-dimensional linkage

3 study techniques overcome this lack of depth perception. These two-dimensional techniques greatly

4 increase the chance that one or more markers used in a study will be close to the sought gene in two-

5 dimensions. This results in more powerful, systematic and efficient methods (including computer

6 programs) and machines for finding genes, such as harmful genes and genes of only modest effect.

7 These techniques also use less dense (more efficient) marker maps (or marker "coverings").

8 The basic principles behind the two-dimensional approach spawn numerous other inventions. These

9 include methods, machines and compositions of matter (groups of molecules) used for gathering the

10 data (i.e. genotype/sample allele frequency data) used in the new two-dimensional studies, and

11 computer techniques for using and handling such data. These techniques work for creatures other than

12 human beings. And they work for markers and genes that are not bi-allelic (any marker or gene can be

13 mathematically transformed to behave like it is bi-allelic). This summary is not exhaustive or limiting,

14 there are other inventions not listed or specifically described here.

15 Brief Description of Drawings

16 Figure 1 is a computer program flowsheet for Process#1 and Process#1A. Figure 1 is a flowsheet

17 illustrating computer programs that execute Process#1 and Process#1A.

18

19 The present patent application is a continuation-in-part of U.S. Patent Application 09/947,768 (filed 5

20 SEPT 2001). And 09/947,768 claims priority from US Provisional 60/230570 (filed 9/5/2000). Patent

21 Application 09/947,768 is a continuation-in-part of U.S. Patent Application 09/623,068 (filed 26 AUG

22 2000). The present patent application is also a continuation-in part of Patent Application 09/623,068

23 (filed 26 AUG 2000). Application 09/623,068 claims priority from PCT/US99/04376 filed (2/26/99).

24 PCT/US99/04376 claims priority from US Provisional applications: 60/076182 filed 27Feb1998,

25 60/086947 filed 27May1998, 60/076102 filed 26 Feb 1998 and 60107673 filed 7 Nov 1998. Each of the

26 following patent applications are incorporated herein by reference in their individual entireties: U.S.

27 Provisional Patent Application 60/230570, PCT/US99/04376, U.S. Patent Application 09/623,068, and

28 U.S. Patent Application 09/947,768.

29 The reader's attention is directed to the following documents or papers each of which is open to the

30 public and each of which is incorporated by reference herein in their entirety: (1) McGinnis, Ewens &

31 Spielman, Genetic Epidemiology 1995 ; 12(6) : 637-40. (2) RE McGinnis Annals of Human Genetics vol

32 62, pp. 159-179.

●   ●

10
Two-Dimensional Linkage Study Methods & Related Inventions

Brief Description of Some Concepts Used By Versions of the Invention

3 Versions of the present invention make use of the novel concept of systematically covering a region on

4 a two-dimensional map similar to an x-y graph with bi-allelic markers. The x axis on this map is the

5 chromosomal location dimension and the y axis of the map is the least common allele frequency

6 dimension. This two-dimensional map is called a CL-F map in this application. (CL stands for

7 chromosomal location and F stands for least common allele frequency.) Each point on a CL-F map has

8 two coordinates: a chromosomal location coordinate and a frequency coordinate. A point on a CL-F

9 map is called a CL-F point.

10 Any one bi-allelic polymorphism (marker or gene) is viewed as being located at a particular CL-F point

11 on a CL-F map. The chromosomal location of the polymorphism is the chromosomal location coordinate

12 of the point. And the least common allele frequency of the polymorphism is the frequency coordinate of

13 the point. The chromosomal location coordinate of a CL-F point is given in units of centiMorgans or

14 base pairs or an equivalent thereof and the least common allele frequency coordinate of a CL-F point is

15 given in units between 0 and 0.5 inclusive, such as 0.2.

16 Distances between any two CL-F points on a CL-F map are given in terms of two numbers:

17 chromosomal location distance and frequency distance. The first number is the distance in the

18 horizontal, chromosomal location direction. This first number is the chromosomal location distance. The

19 second number is the distance in the vertical, frequency direction. This second number is the frequency

20 distance. For example, the CL-F distance δ is given by two numbers δ_{CL} (chromosomal location

21 distance)and δ_F (frequency distance).This is represented as $\delta = [\delta_{CL}, \delta_F]$.

22 The "clustering" of bi-allelic markers near a particular CL-F point is discussed in terms of the number of

23 markers within a particular CL-F distance of the point. For example, if each of N bi-allelic markers is

24 separated from the point by a CL-F distance of less than or equal to δ, then the point is said to be N

25 covered by the markers to within the distance δ. (N being an integer number.)

26 A region on a CL-F map is called a CL-F region. A CL-F region is a collection of one or more CL-F

27 points. Some systematic methods of covering a CL-F region with bi-allelic markers are discussed in

28 terms of the number of markers that are near each point in the region. For example, if each CL-F point

29 in a CL-F region is N covered to within a CL-F distance δ by a subset of a set (or group) of bi-allelic

30 markers, then the region is said to be N covered by the set (or group) of bi-allelic markers to within the

31 distance δ.

32 A set (or group) of bi-allelic markers that cover a CL-F region or a CL-F point is referred to as a set (or

33 group) of bi-allelic covering markers in this application.

34 The inventor discovered that the power of association based linkage tests to detect linkage

35 disequilibrium between a marker and a trait-causing gene (when present) increases greatly when the

36 bi-allelic marker and the bi-allelic gene are located close together on a CL-F map. Systematically

37 covering a CL-F region that is the location of an unknown trait-causing bi-allelic gene with bi-allelic

38 covering markers, therefore greatly increases the power of association based linkage tests to detect

39 linkage disequilibrium (when present) between one or more of the covering markers and the gene.

1 A CL-F matrix is a matrix of rectangular cells of the same length and the same width on a CL-F map.

2 Stipulations that a certain number of covering markers are placed in each cell of the matrix is a method

3 of illustrating particular types of systematic covering of a CL-F region with covering markers.

4 The evidence for linkage obtained from two-dimensional linkage studies is essentially two-dimensional

5 in nature and it is possible to use this two-dimensional information by essentially graphing quantitative

6 evidence for linkage as a function of position in the x-y plane. For example, if quantitative evidence for

7 linkage is represented in the z dimension of a typical three-dimensional x-y-z plot, wherein the x and y

8 dimensions are chromosomal location and least common allele frequency respectively, then it is

9 possible to conceptualize evidence for linkage as occurring in a "hump" or "humps" in the z dimension.

10 And it is possible to analyze the data to find the CL-F location (in the x-y plane) of the peak(s) of this

11 "hump(s)", thus helping to localize a trait-causing gene to the CL-F locale of the peak(s) of the

12 "hump(s)".

13 Versions of the invention also make use of multi-allelic genes and/or markers. It is always possible to

14 combine the alleles of a multi-allelic polymorphism (marker or gene) so that the polymorphism acts

15 mathematically like it is a bi-allelic polymorphism. In effect, it is always possible to mathematically

16 transform a multi-allelic marker or gene to act bi-allelic. Similarly, two or more markers can always be

17 mathematically combined to form a mathematical marker that acts like a single bi-allelic marker. And

18 two or more genes can always be mathematically combined to form a mathematical gene that acts like

19 a single bi-allelic gene. In this application a mathematical bi-allelic marker formed mathematically from

20 one or more markers is called a bi-allelic marker equivalent or BME; and a mathematical bi-allelic gene

21 formed mathematically from one or more genes is called a bi-allelic gene equivalent or BGE.

22 The term true marker or gene is used to distinguish a marker or gene in the ordinary sense from a bi-

23 allelic marker equivalent (BME) or bi-allelic gene equivalent (BGE). The term true allele is used to

24 distinguish an allele in the ordinary sense from a mathematical allele of a BME or BGE. A mathematical

25 allele of a BME or BGE is referred to as an allele equivalent. An allele equivalent is a combination of

26 one or more true alleles or one or more haplotypes.

27 Versions of the invention make use of genes and/or markers, which are not exactly bi-allelic. These

28 genes or markers are approximately bi-allelic. A gene or marker that is approximately bi-allelic almost

29 always occurs in one of two allele forms, however, very rarely it occurs in a different allele form.

30 Various versions of the invention are for genotyping individuals at markers, which systematically cover

31 CL-F regions, or for obtaining sample allele frequency data (such as from pooled DNA) for a sample of

32 individuals for markers which systematically cover CL-F regions. Various versions of the invention are

33 for oligonucleotides used for genotyping individuals at markers which systematically cover CL-F regions

34 or are for obtaining sample allele frequency data (such as from pooled DNA) for a sample of individuals

35 for markers which systematically cover CL-F regions.

36

37 <u>Definitions</u>

38

39 For the purposes of the description and claims the terms used herein will have their generally accepted

40 definition unless otherwise specified.

1   The term **creature** means any organism that is living or was alive at one time. This includes both plants

2   and animals.

3   The term **species** is used in it's broadest sense and includes but is not limited to : 1)biological(genetic)

4   species,2) paleospecies (successional species), 3) taxonomic (morphological ; phenetic) species

5   including species hybrids such as mules, 4) microspecies (agamospecies) 5) biosystematic species(

6   coenospecies,ecosystem species)

7   A **genetic characteristic** is an observable or inferable inherited genetic characteristic or inherited

8   genetic trait including a biochemical or biophysical genetic trait, for example an inherited disease is a

9   genetic characteristic, a predisposition to an inherited disease is a genetic characteristic. A phenotypic

10   characteristic, phenotypic property or character is a genetic characteristic.

11   In this application, **the term gene** means a polymorphism that takes on one or more allele forms and

12   which causes or determines an inherited genetic characteristic or genetic trait. **The term gene** does not

13   mean an entire gene structure with a promoter region, a terminator region, introns, and other parts of an

14   entire gene structure. In this application the term gene means a polymorphism that determines or

15   causes an inherited genetic characteristic and that is part of an entire gene structure in some cases.

16   Each **genetic characteristic** of a creature is **determined** by one or more of the creature's **genes**,

17   wherein the term gene is defined as above.

18   A **segment** is a segment of a chromosome.

19   A **subrange** is a subrange of the least common allele frequency range 0 to 0.5 inclusive.

20   The **width** of a subrange is the difference between the upper and lower limits of the subrange. For

21   example, the width of the subrange 0.1 to 0.4 is $0.4 - 0.1 = 0.3$.

22   **A chromosomal location-least common allele frequency map** is a two-dimensional plot (similar to

23   an x-y graph) wherein the vertical axis(y axis) represents least common allele frequency and the

24   horizontal axis(x axis) represents chromosomal location. A chromosomal location-least common allele

25   frequency map is referred to as a **CL-F map**.

26   **Points on a CL-F map are referred to as CL-F points.** Points on a CL-F map have a chromosomal

27   location coordinate and a least common allele frequency coordinate. CL-F points represent possible

28   chromosomal location and least common allele frequency values for individual bi-allelic markers and

29   genes. Any particular point on a CL-F map is directly opposite a value on the map's least common

30   allele frequency axis(y axis) and is directly opposite a value on the map's chromosomal location axis(x

31   axis). These two values are the two coordinates of the particular point: (1) the chromosomal location

32   coordinate and (2) the least common allele frequency coordinate. A marker or gene located at a

33   particular point on a CL-F map is physically located at the chromosomal location given by the

34   chromosomal location coordinate of the point and the marker or gene's least common allele frequency

35   is the least common allele frequency coordinate of the point. These two coordinates are designated by

36   the term (x, y) wherein x is the value of the chromosomal location coordinate and y is the value of the

37   least common allele frequency coordinate.

38   **A particular CL-F map may be large or small.** For example it is possible for the chromosomal

39   location coordinates of CL-F points on a particular CL-F map to range over an entire chromosome (for

40   example human chromosome number 6). Alternatively it is possible for the chromosomal location

1 coordinates of CL-F points on a particular CL-F map to range over more than one chromosome, for

2 example all the human chromosomes, human chromosomes numbers1 through 22 and X and Y.

3 Similarly it is possible for the chromosomal location coordinates of CL-F points on a particular CL-F

4 map to range over all the chromosomes of a species under study. Alternatively, it is possible for the

5 chromosomal location coordinates of CL-F points on a particular CL-F map to range over a very small

6 segment of chromosome, for example a segment of length 100,000 bp or less. Similarly it is possible for

7 the least common allele frequency coordinates of CL-F points on a particular CL-F map to range over

8 the entire least common allele frequency range 0 to 0.5. Alternatively it is possible for the least common

9 allele frequency coordinates of CL-F points on a particular CL-F map to range over a subrange or

10 subranges of the range 0 to 0.5, for example the subrange 0.1 to 0.2.

11 **If a bi-allelic polymorphism (marker or gene) is said to be located at a particular CL-F point** then

12 the polymorphism's chromosomal location is the chromosomal location coordinate of the point and the

13 polymorphism's least common allele frequency is the least common allele frequency coordinate of the

14 point.

15 **The chromosomal location distance between two CL-F points on a CL-F map** is the absolute

16 difference between the two chromosomal location coordinates of the two points.

17 **The frequency distance between two CL-F points** on a CL-F map is the absolute difference between

18 the two least common allele frequency coordinates of the two points.

19 **The CL-F distance between two CL-F points** is given in terms of two parts or two components : (1)

20 chromosomal location distance and (2) frequency distance. This is denoted as $[D_{CL}, D_F]$, wherein D_{CL} is

21 the chromosomal location distance between the two points and D_F is the frequency distance between

22 the two points. For example [500 bp, 0.3] is an example of a CL-F distance.

23 **If a first CL-F distance is less than or equal to a second CL-F distance** then the chromosomal

24 location distance component of the first CL-F distance is less than or equal to the chromosomal location

25 distance component of the second CL-F distance AND the frequency distance component of the first

26 CL-F distance is less than or equal to the frequency distance component of the second CL-F distance.

27 For example if a first CL-F distance is $[x_1, y_1]$ and a second CL-F distance is $[x_2, y_2]$. And if the first CL-F

28 distance is said to be less than or equal to the second CL-F distance, then x_1 is less than or equal to x_2

29 AND y_1 is less than or equal to y_2.

30 **The term "bi-allelic covering marker(s)" or "covering marker(s)"** is used to distinguish a particular

31 bi-allelic marker or particular bi-allelic markers from other markers. The term is being used simply to

32 avoid ambiguity. In general the term covering marker(s) can be thought of as a marker or markers

33 which have been chosen to cover or serve to cover a CL-F point or a CL-F region.

34 **If a CL-F point is said to be N covered to within a CL-F distance δ by one or more bi-allelic**

35 **covering markers** then the CL-F distance between each of N or more of the covering markers and the

36 point is less than or equal to δ. Wherein N is an integer number greater than or equal to 1.

37 **If a CL-F point is said to be N covered to within a CL-F distance of about (or approximately) δ by**

38 **one or more bi-allelic covering markers** then the CL-F distance between each of N or more of the

39 covering markers and the point is less than or equal to about(or approximately)δ. Wherein N is an

40 integer number greater than or equal to 1.

1   **A CL-F region** is a group of CL-F points. A CL-F region is a region that is or can be represented on a

2   CL-F map. A particular CL-F region may be large or small. For example the chromosomal location

3   coordinates of CL-F points in a particular CL-F region can range over an entire chromosome (for

4   example human chromosome number 6). Alternatively the chromosomal location coordinates of CL-F

5   points in a particular CL-F region can range over more than one chromosome, for example all the

6   human chromosomes, human chromosomes numbers1 through 22 and X and Y. Similarly the

7   chromosomal location coordinates of CL-F points in a particular CL-F region can range over all the

8   chromosomes of a species under study. Alternatively, the chromosomal location coordinates of CL-F

9   points in a particular CL-F region can range over only a small segment of chromosome, for example a

10   segment of length 100,000 bp or less. Similarly the least common allele frequency coordinates of CL-F

11   points in a particular CL-F region can range over the entire least common allele frequency range 0 to

12   0.5. Alternatively the least common allele frequency coordinates of CL-F points in a particular CL-F

13   region can range over only a very small subrange, for example the subrange 0.1 to 0.2 or less.

14   **The length of a CL-F region** is the largest chromosomal location distance between any two CL-F

15   points in the region.

16   **The width of a CL-F region** is the largest frequency distance between any two CL-F points in the

17   region.

18   A **CL-F region that is path connected** is contiguous and it is possible to draw a continuous path

19   between any two points, wherein each point in the path is also in the region.

20   If **a CL-F region is said to be systematically covered by two or more bi-allelic covering markers**

21   then each.point in the region is within a small CL-F distance of one or more of the covering markers,

22   wherein the magnitude of the small CL-F distance is such that there is increased power of an

23   association based linkage test to detect evidence for linkage between one or more covering markers

24   and a gene that is located at a point in the CL-F region, when linkage disequilibrium is present between

25   the gene and one or more of the covering markers.

26   If **a CL-F region is said to be N covered to within a CL-F distance δ by one or more covering**

27   **markers** then each point in the region is N covered to within the CL-F distance δ by the one or more

28   covering markers. Wherein N is an integer greater than or equal to one.

29   If **a CL-F region is said to be N covered to within a CL-F distance of about (or approximately) δ**

30   **by one or more covering markers** then each point in the region is N covered to within the CL-F

31   distance of about(or approximately) δ by the one or more covering markers. Wherein N is an integer

32   greater than or equal to one.

33   **The CL-F distance δ is known as the covering distance** if a CL-F point or CL-F region is N covered

34   to within a CL-F distance δ.

35   **A CL-F covering distance δ has two components:** (1) a chromosomal location distance usually

36   denoted by δ_{CL} and (2) a least common allele frequency distance (abbreviated as frequency distance)

37   usually denoted by δ_F, i.e. $\delta = [\ \delta_{CL},\ \delta_F\ ]$.

38   **The length of a group of covering markers** is determined as follows. The absolute chromosomal

39   location distance between each pair of markers in the group is determined. The greatest absolute

1   chromosomal location distance between each pair of markers in the group is the length of the group of

2   covering markers.

3   **A group of covering markers located on one chromosome can be ordered as a sequence of**

4   **markers** starting with the marker closest to one end of the chromosome and going toward the other

5   end of the chromosome. This is denoted for example as m_1, m_2, m_3, ..., m_{N-2}, m_{N-1}, m_N , wherein N is

6   the number of markers in the group. (The chromosomal location distance between m_1 and m_N is greater

7   than the chromosomal location distance between any other pair of markers in the group and this

8   distance is the length of the group of markers.) **The chromosomal location distance between two**

9   **successive markers in the group** , i.e. between m_R and m_{R+1} , is a **chromosomal intermarker**

10   **distance.** (There are N-1 chromosomal intermarker distances for a group of N covering markers.)

11   **The average chromosomal intermarker distance** for a group is calculated by dividing the length of

12   the group by (N-1), wherein N is the number of covering markers in the group.

13   **The width of a CL-F region** is the largest frequency distance between any two CL-F points in the

14   region.

15   **The length of a CL-F region** is the largest chromosomal location distance between any two CL-F

16   points in the region.

17   A **segment-subrange pair** is the pair formed by pairing a segment of a chromosome and a subrange

18   of the least common allele frequency range 0 to 0.5.

19   The **term segment-subrange** is used as a short version of the term segment-subrange pair. (A

20   segment-subrange is a rectangular region on a CL-F map or a rectangular CL-F region, see below.)

21   If **one or more bi-allelic markers are said to be within(or in) a segment-subrange** then each of the

22   markers is located on (or in) the chromosomal segment of the segment-subrange(pair) and each of the

23   markers' least common allele frequencies is in the subrange of the segment-subrange(pair). (And each

24   of the markers is located within the rectangular region defined by the segment-subrange on a CL-F

25   map.)

26   Alternatively, if **a segment-subrange is said to contain one or more markers or to contain the**

27   **location of one or more markers** then each of the markers is located on (or in) the chromosomal

28   segment of the segment-subrange and each of the markers' least common allele frequencies is in the

29   subrange of the segment-subrange. (And each of the markers is located within or is within the

30   rectangular region on a CL-F map defined by the segment-subrange.)

31   If **one or more CL-F points are said to be within(or in) a segment-subrange** then each of the points

32   is located within the rectangular region defined by the segment-subrange on a CL-F map or on the

33   segment-subrange's borders.

34   **The length of a segment-subrange** is the length of the segment of the segment-subrange.

35   **The width of a segment-subrange** is the width of the subrange of the segment-subrange.

36   **The area of a segment-subrange** is the segment subrange's length multiplied by the segment

37   subrange's width.

38   If **a CL-F region is said to comprise a segment-subrange**, then each point in the segment-subrange

39   is in(or included in) the CL-F region.

1   **If a CL-F region is said to comprise an area of greater than or equal to X multiplied by Y**, then the

2   CL-F region comprises one or more nonoverlapping segment-subranges, and the sum of the areas of

3   the segment-subranges is greater than or equal to X multiplied by Y.

4   A **CL-F matrix** is a collection of segment-subranges, wherein each segment-subrange of the collection

5   has the same width and the same length. Each segment-subrange in the collection (or the matrix) is a

6   **CL-F matrix cell.** Any one CL-F matrix cell in a CL-F matrix shares two or more of the cell's borders

7   with two or more other cells in the matrix. And all the cells in a CL-F matrix together form a single

8   segment-subrange. A CL-F matrix is characterized by the length and the width of the cells in the matrix

9   denoted by length x width, or $L_{MC} x W_{MC}$, wherein L_{MC} is the length of each cell in the matrix and W_{MC} is

10   the width of each cell in the matrix. A CL-F matrix is also characterized by the number of rows of cells,

11   R_M, in the matrix. And a CL-F matrix is characterized by the number of columns of cells, C_M, in the

12   matrix. There are two or more cells in a CL-F matrix. A CL-F matrix is also characterized by the point of

13   origin of the matrix, denoted by (cl_O, f_O). The point of origin of a CL-F matrix is at any chromosomal

14   location and cl_O takes on any reasonable value in an entire species genome. The point of origin of a

15   CL-F matrix is at any one value in the least common allele frequency range 0 to 0.5. (A CL-F matrix is

16   similar to the squares of a chessboard or to equal rectangular floor tiles that are all oriented in the same

17   direction and cover a rectangular floor. One corner of the matrix is the matrix's point of origin.)

18   The **width of each cell of a particular CL-F matrix** is any value greater than zero and less than 0.5.

19   The width of a cell is often denoted by W_{MC}.

20   Any length in chromosomal location distance units is chosen for the **length of each cell of a particular**

21   **CL-F matrix**. The length of a cell is often denoted by L_{MC}.

22   The **centerpoint** of a CL-F matrix cell is in the center of the cell. The centerpoints of a CL-F matrix form

23   a **matrix centerpoint lattice.** Each point of a matrix centerpoint lattice is separated by a CL-F distance

24   of $[0, W_{MC}]$ or $[L_{MC}, 0]$ from two or more neighboring centerpoints.

25   If **one or more bi-allelic markers are in(or within) the segment-subrange that is a CL-F matrix**

26   **cell,** then each of the markers is in or within the CL-F matrix cell.

27   If **one or more CL-F points is in (or within) a CL-F matrix**, then each of the points is in or within a cell

28   of the matrix.

29   If **a CL-F region comprises a CL-F matrix**, then each point that is in the matrix is also in the region.

30   If **a CL-F region is a CL-F matrix**, then the region consists of the points that are in the matrix.

31   If two CL-F matrix cells share a common border, then the **two CL-F matrix cells are in contact.**

32   If two CL-F matrix cells share a common corner, then the **two CL-F matrix cells are touching.** (Two

33   cells that are in contact are also touching.)

34   If **a group of CL-F points is connected to within a CL-F distance [X,Y]**, then for any two points in

35   the group, denoted p_1 and p_R, there is an ordered sequence of points in the group denoted p_1, p_2,

36   p_3,......, p_{R-2}, p_{R-1}, p_R, R being an integer greater than or equal to 2, wherein the CL-F distance between

37   each point in the sequence and the next point in the sequence is less than or equal to [X,Y]. **The**

38   **distance [X,Y] is the connecting distance.** (Put in simple terms if a group of points is connected to

39   within [X,Y], then there is a path between each pair of points in the group, the path consisting of a

40   series of steps, wherein each step in the path is a movement between two points in the group that are

1   separated by a CL-F distance of less than or equal to [X,Y]. A simple group of points connected to

2   within a CL-F distance of [X,Y] is a group of three points, wherein each point in the group is within a CL-

3   F distance of less than or equal to [X,Y] of another point in the group. The concept of connectivity

4   introduced here is similar to the basic concept of connectivity in mathematical graph theory.)

5   If **a group of N markers is connected to within a CL-F distance [X,Y]**, wherein N is an integer, then

6   each of the markers is located at one point of a group of N points, the group of N points being

7   connected to within a CL-F distance [X,Y].

8   If **two bi-allelic markers are said to be in extreme positive disequilibrium** then d is approximately

9   equal to d_{max} for the two markers, which for the purposes of this definition are designated marker M

10   with least common allele A and marker m with least common allele B. Wherein according to standard

11   usage, the disequilibrium coefficient (d) is defined by the equation d=f(AB) - f(A)f(B) where f(A) and f(B)

12   are defined as the population frequencies of alleles A and B, respectively, and f(AB) is the population

13   frequency of the AB haplotype. And d_{max} is defined as the maximum possible positive value of d

14   assuming the allele frequencies of A and B are f(A) and f(B), and thus d_{max}= q-f(A)f(B) where q is the

15   lesser of f(A) and f(B). (In this application d is used to represent the disequilibrium coefficient; the

16   symbol δ is often used in scientific papers to represent the disequilibrium coefficient.)

17   If **a pair of markers is said to be in extreme positive disequilibrium**, then the two markers of the

18   pair are in extreme positive disequilibrium.

19   If **a pair of bi-allelic markers is said to be redundant within distance D** then the two markers of the

20   pair are in extreme positive disequilibrium and the two markers are located on the same chromosome

21   and the two markers are located within a CL-F distance D of each other on a CL-F map, wherein D is a

22   specified distance and D has two components, a chromosomal location distance component D_{CL} and a

23   frequency distance component, D_F; D = [D_{CL}, D_F].

24   An **allele equivalent (AE)** is a group of one or more "haplotype values" of one or more polymorphisms

25   of the same type, either markers or genes. (For the purposes of this application a haplotype value of

26   one polymorphism is equivalent to an allele value at the one polymorphism.)The group of haplotype

27   values is then analyzed as if the group is a single allele at a bi-allelic polymorphism; the group of

28   haplotype values acts as a single allele at a bi-allelic polymorphism; the collection of the one or more

29   polymorphisms upon which the haplotype values are based acts as a bi-allelic polymorphism; the

30   collection of one or more polymorphisms forms a bi-allelic **polymorphism equivalent (PE)** that acts as

31   a bi-allelic polymorphism; **the polymorphism equivalent has(or possesses) the allele equivalent.**

32   The allele equivalent belongs to the polymorphism equivalent. In this application, **each polymorphism**

33   **equivalent is a bi-allelic marker equivalent(BME) or a bi-allelic gene equivalent(BGE).**

34   **A bi-allelic marker equivalent (BME)** is one or more markers and a grouping of the haplotype values

35   of the one or more markers into two groups (e.g. group I and group II)(For the purposes of this

36   application a "haplotype value" of one marker is equivalent to an allele at the one marker). The one or

37   more markers and the two groups of haplotype values of the one or more markers are then analyzed as

38   if the one or more markers are a single bi-allelic marker with alleles I and II. Each group of the groups I

39   and II is an allele equivalent. For example, a multi-allelic microsatellite marker has it's multiple alleles

40   grouped into two groups and the microsatellite marker and these two groups of alleles then act

1   equivalent to a bi-allelic marker and are analyzed as if the microsatellite marker with the two groups is

2   bi-allelic (for an example of this see McGinnis, Ewens & Spielman, Genetic Epidemiology 1995 ; 12(6) :

3   637-40, which is incorporated herein by reference)

4   Also for example, two or more multi-allelic markers have their haplotypes separated into two groups of

5   haplotypes and the multi-allelic markers with their two groups of haplotypes are analyzed as if they

6   were a single bi-allelic marker.

7   For example bi-allelic marker A has alleles a and a* and bi-allelic marker B has alleles b and b*. Then

8   the four haplotypes ab, ab*, a*b* and a*b are grouped into two groups, for example group I: ab and a*b*

9   and group II : ab* and a*b. Then a BME formed by markers A and B takes on values of group I (or I) for

10   haplotypes ab or a*b* or group II (or II) for the haplotypes ab* or a*b ; and the two markers and the two

11   group values(I and II) are analyzed as though they form a single bi-allelic marker(the BME). The same

12   type of reasoning and procedure is extended to 3 or more bi-allelic markers, 3 or more bi-allelic marker

13   equivalents or 2 or more multi-allelic markers.

14   (Logically, of course, the genotype at a BME for an individual is determined by knowing the two

15   haplotype values at the one or more markers that form the BME for each of the individual's two

16   homologous chromosomes that carry the one or more markers. The genotype is then determined by

17   classifying each haplotype as belonging to group I or group II or the equivalent thereof. The three

18   possible genotype values at the BME are I / I, I / II, and II / II or the equivalent thereof.)

19   Similarly, **a bi-allelic gene equivalent (BGE)** is one or more genes and a grouping of all the haplotype

20   values of the one or more genes into two groups (e.g. group I and group II).

21   For the purposes of the description and claims, **the chromosomal location of a polymorphism**

22   **equivalent** is at any point on the smallest chromosomal segment that contains the one or more

23   polymorphisms that form the polymorphism equivalent(PE).

24   **The allele frequency of an allele equivalent (AE)** is determined as follows. An allele equivalent (AE)

25   is a group of haplotype values of one or more polymorphisms. The frequency of the allele equivalent is

26   the sum of the frequencies of the haplotype values in the group that makes up the allele equivalent.

27   For the purposes of the application, description, claims and definitions the term **true allele** is used to

28   distinguish an allele according to standard usage (i.e. at a single polymorphism) from an allele

29   equivalent (AE).

30   **The least common allele frequency of a bi-allelic polymorphism equivalent (BPE)** is determined

31   as follows. Each of the two groups(I and II) of the haplotype values of the one or more polymorphisms

32   which form the BPE is assigned a frequency. The frequency of I is the sum of the frequencies of the

33   haplotype values in group I. And the frequency of II is the sum of the frequencies of the haplotype

34   values in group II. The least of the frequency of I and the frequency of II is the least common allele

35   frequency of the BPE. If the frequency of I and the frequency of II are equal, then the least common

36   allele frequency of the BPE is the frequency of I or the frequency of II.

37   For the purposes of the description and claims, **the chromosomal location of a bi-allelic marker**

38   **equivalent (BME)** is at any point on the smallest chromosomal segment which contains the one or

39   more markers which form the BME.

1   **The chromosomal location distance from a BME to a CL-F point** on a CL-F map is the shortest

2   chromosomal location distance from the CL-F point to any one of the one or more markers which form

3   the BME.

4   **The least common allele frequency of a bi-allelic marker equivalent (BME)** is determined as

5   follows. Each of the two groups(I and II) of the haplotype values of the one or more markers which form

6   the BME is assigned a frequency. The frequency of I is the sum of the frequencies of the haplotype

7   values in group I. And the frequency of II is the sum of the frequencies of the haplotype values in group

8   II. The least of the frequency of I and the frequency of II is the least common allele frequency of the

9   BME. If the frequency of I and the frequency of II are equal, then the least common allele frequency of

10   the BME is the frequency of I or the frequency of II.

11   **The frequency distance from a BME to a CL-F point** on a CL-F map is the absolute difference

12   between the least common allele frequency of the BME and the least common allele frequency

13   coordinate of the CL-F point.

14   (If a CL-F point on a CL-F map is covered by one or more BMEs to within a distance δ, wherein $\delta = [\delta_{CL}$

15   , δ_F], then the CL-F distance from each of the one or more BMEs to the CL-F point is less than or equal

16   to δ. And the chromosomal location distance from one of the markers which form each BME to the CL-F

17   point is less than or equal to δ_{CL} . And the frequency distance from each of the one or more BMEs to the

18   CL-F point is less than or equal to δ_F .)

19   **A bi-allelic marker equivalent is in(or within) each CL-F matrix cell that contains the**

20   **chromosomal location of the bi-allelic marker equivalent (BME).** (Since the chromosomal location

21   of a bi-allelic marker equivalent (BME) is at any point on the smallest chromosomal segment which

22   contains the one or more markers which form the BME, in some cases, a bi-allelic marker equivalent is

23   in more than one CL-F matrix cell.)

24   For the purposes of the application, the term **true bi-allelic marker** is used to distinguish a bi-allelic

25   marker with two alleles according to usual usage (i.e. at a single polymorphism) from a bi-allelic marker

26   equivalent(BME). A true bi-allelic marker is not a bi-allelic marker equivalent (BME). The term **true bi-**

27   **allelic polymorphism** is used to distinguish a bi-allelic polymorphism with two alleles according to

28   usual usage from a bi-allelic polymorphism equivalent(BPE).

29   The term **true allele** of a true bi-allelic marker means an allele of a true bi-allelic marker.

30   **A polymorphism(marker or gene) which is exactly bi-allelic** has exactly two alleles and the sum of

31   the frequency of each of the two alleles is 1; for example if the two alleles are A and B, then f(A) + f(B)

32   = 1. A polymorphism that is exactly bi-allelic is a true bi-allelic polymorphism with exactly two true

33   alleles or a bi-allelic polymorphism equivalent (BPE) with exactly two allele equivalents.

34   **A polymorphism(marker or gene) which is approximately bi-allelic** has three or more alleles. And

35   the polymorphism has a first allele and a second allele; and the sum of the frequency of the first allele

36   and the frequency of the second allele is approximately 1. And the frequency of the first allele and the

37   frequency of the second allele is much greater than the sum of the allele frequencies of all the alleles of

38   the polymorphism that are not the first or the second alleles. For the versions of the invention for bi-

39   allelic polymorphisms (bi-allelic markers and bi-allelic genes) described herein, a polymorphism which

40   is approximately bi-allelic is analyzed as if the polymorphism has only two alleles, the first allele and the

1   second allele. For the versions of the invention described herein, the least common allele frequency of

2   a polymorphism which is approximately bi-allelic, is the least of the frequencies of the first and the

3   second alleles of the polymorphism. A polymorphism which is approximately bi-allelic is a true

4   polymorphism with true alleles (the allele frequencies of the true alleles conform to the stipulations of

5   this definition) or is a bi-allelic polymorphism equivalent(BPE) with allele equivalents(the allele

6   frequencies of the allele equivalents conform to the stipulations of this definition).

7   **SNP** stands for single nucleotide polymorphism.

8   **A statistical linkage test based on allelic association** is any mathematical test, mathematical

9   computation or equivalent thereof which gives a quantitative estimate (or equivalent thereof) of

10   evidence for linkage of a polymorphic marker and phenotypic trait (genetic characteristic) based on

11   association between one or more of the alleles of the marker and the phenotypic trait in a sample of

12   individuals of a population of a species. A statistical linkage test based on allelic association is any

13   statistical test that detects or suggests linkage on the basis of allelic association. A statistical linkage

14   test based on allelic association includes tests which suggest but do not prove linkage such as

15   comparison of marker allele frequencies in disease cases and in unrelated controls. A statistical linkage

16   test based on allelic association is also any test such as the TDT which may be regarded as "proving"

17   linkage. (A statistical linkage test based on allelic association can, of course, give an estimate of the

18   association of one or more allele equivalents of a marker equivalent and a genetic characteristic; see

19   definition of BME above.) One aspect of a statistical linkage test based on allelic association is it's

20   potential use to calculate the probability, or equivalent thereof, that there is genuine association of one

21   or more of the alleles of the marker and a genetic characteristic for the population as a whole (rather

22   than just for the sample alone). A statistical linkage test based on allelic association is an association

23   based linkage test. (The term **population** in this application is used in a statistical sense and means a

24   group of individuals. The term **population** in this application is not used purely in the sense the term

25   **population** is used in the field of population genetics.)

26   As noted in this application, in PCT/US99/04376 and the paper Annals of Human Genetics (1998), 62,

27   159-179, P_t and P_s are binomial probabilities that determine the magnitudes of the TDT chi-square

28   statistic and ASP test statistic respectively. Both P_s and P_t increase substantially as the frequencies of

29   a disease-causing allele and positively associated marker allele become similar in magnitude. (see

30   Theory of Operation Section PCT/US99/04376 and here). Thus this application teaches the utility of

31   using both a TDT chi-square statistic (or similar statistical test) and an ASP test statistic (or similar

32   statistical test) **together in combination.** For the purposes of this application then, a statistical linkage

33   test based on allelic association also includes tests having the characteristics described in the

34   paragraph immediately above that are combined with an ASP test statistic (or similar statistic).

35   The term **sample** means a group of individuals which is a subset of a population.

36   In this application, **an allele is considered to be a piece of double stranded DNA** that is singular or

37   distinctive for the allele. The piece of double stranded DNA that is distinctive for the allele contains the

38   particular DNA sequence that distinguishes the allele from other alleles (alternate sequences) at the

39   polymorphic site of interest plus two double stranded "flanking" DNA sequences, one flanking DNA

1 sequence being on one side of the polymorphic site and the other flanking DNA sequence being on the
2 other side of the polymorphic site.
3 **Alternate strand of an allele:** A double stranded piece of DNA that is distinctive for an allele consists
4 of two pieces of single stranded DNA which are exactly complementary to one another. The two pieces
5 of single stranded DNA are referred to as the two strands of the allele. Each of the two strands of the
6 allele is the alternate of the other strand of the allele. For the purposes of this definition, the two strands
7 are referred to as the first strand and the second strand. The alternate strand of the first strand is the
8 second strand. And the alternate strand of the second strand is the first strand. Each strand of an allele
9 is exactly complementary to the strand's alternate strand.
10 **An oligonucleotide** is either a single or double stranded oligonucleotide. The length of an
11 oligonucleotide ranges from a few bases or base pairs to approximately any number of bases or base
12 pairs in the DNA sequence of any allele.

13 **An oligonucleotide, either single or double stranded, is complementary** to an allele if the DNA
14 sequence of each strand of the oligonucleotide is exactly or approximately complementary to all or part
15 of the DNA sequence of one of the DNA strands of the allele and the oligonucleotide has utility in
16 identifying the allele by a hybridization reaction or equivalent thereof similar to as described below
17 under oligonucleotide technology.

18 **An allele is identified by a hybridization reaction with an oligonucleotide that is complementary**
19 **to the allele.** In this application **there are two types of oligonucleotides that are complementary to**
20 **an allele.** The two types of oligonucleotides complementary to an allele are identified as type(1) or
21 type(2).

22 A type (1) complementary oligonucleotide is complementary to the part of an allele's DNA sequence
23 that actually contains the allele's polymorphic site; and the type(1) complementary oligonucleotide has
24 utility to identify the allele by means of a hybridization reaction of the oligonucleotide to the part of the
25 allele's DNA sequence that actually contains the allele's polymorphic site. A hybridization reaction of a
26 type(1) oligonucleotide to the part of an allele's DNA sequence that actually contains the allele's
27 polymorphic site is a type (1) hybridization reaction.

28 A type (2) complementary oligonucleotide is complementary to an allele at a DNA sequence that flanks
29 (but does not contain) the allele's polymorphic site; and the type (2) complementary oligonucleotide has
30 utility to identify the allele by means of a hybridization reaction wherein the oligonucleotide hybridizes to
31 the allele at a DNA sequence that flanks (but does not contain) the allele's polymorphic site and
32 identification of the allele is subsequently achieved by extension of the oligonucleotide (and possibly
33 one or more other type(2)complementary oligonucleotides) across the polymorphic site with a DNA
34 polymerase such as occurs, for example, in a standard PCR (polymerase chain reaction). A
35 hybridization reaction of a type(2) oligonucleotide to an allele at a DNA sequence that flanks (but does
36 not contain) the allele's polymorphic site is a type (2) hybridization reaction.
37 Each version of **oligonucleotide technology** is a means to test for the presence (or absence) of each
38 of one or more true alleles of a group of true alleles in an individual's chromosomal DNA. The presence

or absence of any one true allele in the group is tested for by means of a type (1) or type (2)

hybridization reaction (or equivalent) with an oligonucleotide that is complementary(type(1) or type(2))

to the true allele. Put another way, the presence or absence of each true allele in the group is tested for

by means of a type(1) or type(2)hybridization reaction (or equivalent) with an oligonucleotide that is

complementary to each true allele in the group. There are many versions of oligonucleotide technology,

some of these versions are described in more detail below. (In this application, the term "chromosomal

DNA" includes chromosomal DNA obtained directly from an individual as well as DNA obtained as

amplification products using PCR and chromosomal DNA obtained directly from an individual.

A physico-chemical signal is any physical (including chemical) signal which is detected by human

senses or by apparatus. A physico-chemical signal includes, but is not limited to, (1) an electrical signal

such as is generated when oligonucleotides that are attached to a silicon chip hybridize with

complementary alleles, (2) a visual or optical signal such as is generated when oligonucleotides

attached to a glass slide hybridize with complementary alleles, (3) a signal (such as a dye color)

generated by the products of a PCR(polymerase chain reaction) such as when oligonucleotides that are

used as primers for PCR reactions hybridize with complementary alleles.

The collection of true alleles of a group of one or more bi-allelic markers is defined as consisting

of each true allele of each true marker in the group and each true allele of each haplotype that forms

each allele equivalent of each marker equivalent in the group.

If **a set of oligonucleotides is said to be complementary to a group of one or more bi-allelic**

markers, then each oligonucleotide in the set is type(1) or type(2) complementary to at least one of the

true alleles in the collection of true alleles of the group of one or more markers; and there is an

oligonucleotide in the set that is type(1) or type(2) complementary to each true allele in the collection of

true alleles of the group of one or more markers.

Sample allele frequency data for a marker and a sample is obtained by pooling DNA specimens

from individuals of the sample into one or more DNA pools. An allele frequency for each of the marker's

alleles is obtained for each DNA pool. In the case of a bi-allelic marker, determining the sample allele

frequency for one allele essentially determines the sample allele frequency for the other allele. (For

example, in some association based linkage studies, each DNA pool contains DNA from individuals of

the sample with the same or similar phenotype status.) (It is also possible to obtain sample allele

frequency for a marker and a sample by calculation using genotype data at the marker for each

individual in the sample.)

Genotype data/sample allele frequency data for a marker and a sample is (1)genotype data at the

marker for each individual of the sample, or (2)a combination of genotype data at the marker for one or

more individuals in the sample and sample allele frequency data for the marker for the sample, or

(3)sample allele frequency data for the marker for the sample. In the case of genotype data, DNA

specimens from individuals are tested individually to determine genotype. In the case of sample allele

frequency data DNA specimens from individuals are pooled, or sample allele frequency is calculated

using genotype data for each individual in the sample.

Description

For the versions of the invention described herein and the claims, **a bi-allelic genetic characteristic gene or a bi-allelic gene** is a gene which is exactly bi-allelic or a gene which is approximately bi-allelic. For the versions of the invention described herein and the claims, **a bi-allelic genetic characteristic gene or a bi-allelic gene** is a gene which is a true bi-allelic gene or a bi-allelic gene equivalent (BGE). A bi-allelic gene equivalent is exactly bi-allelic or approximately bi-allelic. A true bi-allelic gene is exactly bi-allelic or approximately bi-allelic.

For the versions of the invention described herein and the claims, **a bi-allelic marker or a bi-allelic covering marker** is a marker which is exactly bi-allelic or a marker which is approximately bi-allelic. Each marker that is exactly bi-allelic is a true bi-allelic marker or a bi-allelic marker equivalent. And each marker that is approximately bi-allelic is a true bi-allelic marker or a bi-allelic marker equivalent (BME).

Process #1, A process for identifying one or more bi-allelic markers linked to a bi-allelic genetic characteristic gene in a species of creatures, comprising:

a)choosing two or more bi-allelic covering markers so that a CL-F region is systematically covered by the two or more covering markers;

b)choosing a statistical linkage test based on allelic association for each covering marker;

c)choosing a sample of individuals for each covering marker ;

d)obtaining genotype data/sample allele frequency data for each covering marker and the sample chosen for each covering marker, and obtaining phenotype status data for the genetic characteristic for each individual in the sample chosen for each covering marker;

e)calculating evidence for linkage between each covering marker and the gene using the statistical linkage test based on allelic association chosen for each covering marker and the genotype data/sample allele frequency data for each covering marker and using the phenotype status data for the genetic characteristic for each individual in the sample chosen for each covering marker obtained in d); and

f)identifying those covering markers as linked to the genetic characteristic gene which show evidence for linkage based on the calculations of e).

The following is a more detailed description of process #1.

1

2 **Process #1, A process for identifying one or more bi-allelic markers linked to a bi-allelic genetic**

3 **characteristic gene in a species of creatures comprising:**

4

5 **a)choosing two or more bi-allelic covering markers so that a CL-F region is systematically**

6 **covered by the two or more covering markers;** Any method of systematically covering the CL-F

7 region is acceptable. In this application, the systematic covering of a CL-F region in versions of the

8 invention is described mathematically as the covering of a CL-F region, wherein the CL-F region is N

9 covered to within a CL-F distance δ by two or more bi-allelic covering markers. For further details

10 regarding a), see <u>Detailed Description of the Systematic Covering of a CL-F Region Used In Versions</u>

11 <u>of the Invention</u> below.

12

13 **b)choosing a statistical linkage test based on allelic association for each covering marker ;** The

14 statistical linkage test based on allelic association chosen for any one particular covering marker is any

15 statistical linkage test based on allelic association as defined in the definitions section. Statistical

16 linkage tests based on allelic association are described in the genetics and population genetics

17 literature and are known to those of ordinary skill in the art. Some examples of a statistical linkage test

18 based on allelic association are the TDT, Haplotype Relative Risk Method(HRR) and Allele Frequency

19 Comparison In Disease Cases Versus Unrelated Controls .It is possible for different statistical linkage

20 tests based on allelic association to be chosen for different covering markers. For purposes of technical

21 convenience, the same statistical linkage test based on allelic association is preferably chosen for each

22 covering marker.

23

24 **c)choosing a sample of individuals from the species for each covering marker ;** For the process

25 to be workable, the sample chosen for any one covering marker must be suitable for the statistical

26 linkage test of b) above chosen for the covering marker. Knowledge of a suitable sample for the

27 statistical linkage test chosen in b) above for the covering marker is within the understanding of a

28 person skilled in the art. For purposes of technical convenience, the same sample of individuals is

29 preferably chosen for each covering marker.

30

31 **d)obtaining genotype data/sample allele frequency data for each covering marker and the**

32 **sample chosen for each covering marker, and obtaining phenotype status data for the genetic**

33 **characteristic for each individual in the sample chosen for each covering marker;**

34 Sample allele frequency data for any one covering marker for the sample chosen for the covering

35 marker is obtained by pooling DNA from individuals of the sample into one or more DNA pools. It is also

36 possible to obtain sample allele frequency data for any one covering marker by calculation using

37 genotype data at the marker for each individual in the sample. Each DNA pool contains DNA from

38 individuals of the sample with the same or similar phenotype status. An allele frequency for each of the

39 marker's alleles is obtained for each pool. Genotype data/sample allele frequency data for any one

40 covering marker is (1)genotype data at the covering marker for each individual in the sample chosen for

1 the covering marker, or (2)a combination of genotype data at the covering marker for one or more

2 individuals in the sample chosen for the covering marker and sample allele frequency data for the

3 covering marker for the sample chosen for the covering marker, or (3)sample allele frequency data for

4 the covering marker for the sample chosen for the covering marker. The genotype data/sample allele

5 frequency data for any one covering marker must be suitable for the statistical linkage test based on

6 allelic association chosen for the covering marker in b). It is possible to choose different types of

7 genotype data/sample allele frequency data for each covering marker. For purposes of technical

8 convenience, the same type of genotype data/sample allele frequency data (1), (2), or (3) is chosen for

9 each covering marker. Some examples of ways to practice d) are the use of technology cited under

10 Oligonucleotide Technology (below) or mass spectrometry (such as MALDITOF).[1]

11

12 **e)calculating evidence for linkage between each covering marker and the gene using the**

13 **statistical linkage test based on allelic association chosen for each covering marker and the**

14 **genotype data/sample allele frequency data for each covering marker and using the phenotype**

15 **status data for the genetic characteristic for each individual in the sample chosen for each**

16 **covering marker obtained in d); and**

17

18

19 **f)identifying those covering markers as linked to the gene which show evidence for linkage**

20 **based on the calculations of e).**

21 The meanings of d), e) and f) are within the understanding of those of ordinary skill in the art. Fine

22 points of using a statistical linkage test based on allelic association as a measure of evidence for

23 linkage are known to those in the art.[2]

24

25 Process #1 described above is equivalent to localizing a genetic characteristic gene to a particular

26 chromosomal location (i.e. a sub-region of a particular chromosome.) This is because markers which

27 are linked to a gene are also physically close to the gene in terms of physical (chromosomal) location.

28 To locate a gene causing the genetic characteristic of Process #1, the gene is localized to the

29 approximate chromosomal location of one or more covering markers which are identified as showing

30 evidence for linkage in f).

31 **Process#1A** It is also possible to use Process #1 to localize a genetic characteristic gene to an

32 approximate CL-F location(chromosomal location-least common allele frequency location). Such a

33 process is expressed as follows:

34 **Process#1A : A process for localizing a bi-allelic genetic characteristic gene in a species of**

35 **creatures to a chromosomal location-least common allele frequency (CL-F) location, comprising**

36 **the a), b), c), d) and e) of Process #1 and further comprising:**

37 **f)localizing the gene to the chromosomal location-least common allele frequency (CL-F) location**

38 **of one or more markers that show evidence for linkage based on the calculations of e).**

39 It is the teaching of this application that the strength of evidence for linkage increases as markers that

40 are in linkage disequilibrium with a gene become close to the gene on a CL-F map. It is possible for f)

1 to be done by an individual plotting data by hand and examining the data. It is also possible for software

2 to perform f). It is possible for f) to include using the dependence of quantitative evidence for linkage of

3 e) on CL-F location. For example, if quantitative evidence for linkage calculated in e) (of process #1 or

4 #1A) is represented in the z dimension of a typical three-dimensional x-y-z plot, wherein the x and y

5 dimensions are chromosomal location and least common allele frequency respectively, then it is

6 possible to conceptualize evidence for linkage as occurring in a "hump" (or "humps")in the z dimension.

7 And it is possible to use the evidence for linkage calculated in e) of (process #1 or #1A) to find the CL-F

8 location (in the x-y plane) of the peak(s) of a "hump(s)", thus helping to localize a trait causing gene to

9 the CL-F locale of the peak(s) of the "hump(s)". For example it is possible to use computer

10 programming techniques that detect gradients such as, for example, linear or nonlinear programming

11 techniques in mathematical optimization theory[3] to find the peak(s) of a hump(s) in this (Process #1A

12 described above is equivalent to localizing a genetic characteristic gene to a particular chromosomal

13 location (i.e. a sub-region of a particular chromosome.) This is because localizing a gene to a particular

14 CL-F region also localizes the gene to a particular chromosomal region.)

15 **Software**

16 A computer program that executes each step or step-like part of Process#1 is an example of

17 Process#1. A computer program that executes each step or step-like part of Process#1A is an example

18 of Process#1A. A flowsheet illustrating programs that execute Process#1 and Process#1A is entitled

19 Drawing #1(see drawing section). It is also possible for a computer program to execute any one of(or

20 one or more combinations of) the steps or step-like parts of Process#1 or Process#1A. A person of

21 ordinary skill in the art could write such a program without undue experimentation. The level of skill at

22 computer programming in the art is great as evidenced by numerous computer programs. Some

23 computer programs in the art are programs such as MAPMAKER/SIBS[4] , GENEHUNTER[5] , LINKAGE[6]

24 , and FASTLINK.[7]

25 <u>Detailed Description of the Systematic Covering of a CL-F Region Used In Versions of the Invention</u>

26 (see definitions section for meaning of CL-F region that is systematically covered). The CL-F region and

27 covering markers are for a species and the one or more individuals are members of the species. The

28 chromosomal location coordinate of each covering marker is based on information regarding the

29 chromosomal location of each covering marker. One such source of information is chromosomal maps.

30 Chromosomal maps are provided by such institutions as the Whitehead Institute or Marshfield

31 Foundation for Biomedical Research. Chromosomal maps include, but are not limited to genetic maps,

32 physical maps, and radiation hybrid maps.

33 The least common allele frequency coordinate of each covering marker is based on any reasonable

34 information regarding the least common allele frequency of each covering marker. It is possible to use

35 information from different populations for the allele frequencies of different covering markers. For

36 example, it is possible for the least common allele frequencies of two different covering markers to be

37 based on information from two different, but similar populations. For purposes of technical convenience,

38 the least common allele frequency of each covering marker is based on information from the same

39 population. One source of information on least common allele frequency is institutions which provide

40 chromosomal maps such as the Whitehead Institute or Marshfield Foundation for Biomedical Research.

Systematic Covering Of A CL-F Region, Wherein A CL-F Region Is N Covered To Within A CL-F Distance δ By Two or more Bi-Allelic Covering Markers

In this application, the systematic covering of a CL-F region in versions of the invention is described mathematically as the covering of a CL-F region, wherein the CL-F region is N covered to within a CL-F distance δ by two or more bi-allelic covering markers. The covering markers are chosen so that the CL-F region is N covered to within the CL-F distance δ by using information regarding the chromosomal location and least common allele frequency of each covering marker.

It is possible for the chromosomal location component of δ to be as great as about any chromosomal length, computed by any method, for which linkage disequilibrium has been observed between any polymorphisms in any population of the species. It is preferable in terms of increasing the power of a version of the invention for linkage studies that the chromosomal location component of δ be less than about the greatest chromosomal length, computed by any method, for which linkage disequilibrium has been observed between any polymorphisms in any population of the species. In general, the smaller the chromosomal location component of δ, the greater the power of a version of the invention for linkage studies.

It is possible for the frequency distance component of δ to be as great as about 0.2. (Depending on the penetrance ratio (r) or the disequilibrium between marker and gene, it is also possible for the frequency distance component of δ to be greater than 0.2 under some conditions as evidenced by Table 2 under Theory of Operation. So it is also possible for the frequency distance component of δ to be as great as about 0.25 or higher.) It is preferable in terms of increasing the power of a version of the invention for linkage studies that the frequency distance component of δ to be less than about 0.2. In general, the smaller the frequency distance component of δ, the greater the power of a version of the invention for linkage studies.

Linkage disequilibrium has been observed between polymorphisms separated by 10 to 12 cM in some homogeneous human populations. Therefore, it is possible for the chromosomal location distance component of δ to be as large as about 10 to 12 cM, about 10 to 12 million bp, or the equivalent thereof for homogeneous human populations. It is preferable in terms of increasing the power of a version of the invention for linkage studies in human populations that δ is less than or equal to about [1 million bp, 0.15] or the equivalent thereof. It is more preferable in terms of increasing the power of a version of the invention for linkage studies in human populations that δ is less than or equal to about [250,000 bp, 0.1] or the equivalent thereof.

In general, the smaller the magnitude of δ is in terms of either frequency distance, chromosomal location distance, or both, the greater the power of a version of the invention for linkage studies. In general, the greater N is, the greater the power of a version of the invention for linkage studies. Because the greater N is, the greater the chance that linkage is detected between one or more covering markers and a gene or genes. The largest that N is chosen is limited by the number of known markers in the neighborhood of the CL-F region and also by the distribution of the known markers.

In general, the larger the CL-F region which is N covered, the greater the power of a version of the invention for linkage studies, because a larger region is scanned (covered). Less dense coverings

1 wherein N is small, and the magnitude of δ is large also have technical and economic advantages for

2 certain situations.

3 **Specific types of CL-F regions that are N covered**

4 Specific types of CL-F regions that are N covered are useful. For example, a rectangular CL-F region, a

5 segment-subrange, that is N covered is used in an association based linkage study to test for the

6 presence of a trait causing bi-allelic gene located within the segment-subrange. In the case in which a

7 group of points is N covered to within a CL-F distance [x,y] and the group of points is connected to

8 within a CL-F distance of [2x,2y] or less, then a path connected CL-F region is N covered to within the

9 CL-F distance [x,y].

10 A CL-F matrix is a device to illustrate and describe the systematic nature of special cases of CL-F

11 regions that are N covered. In the case in which there are N or more markers within each cell of a CL-F

12 matrix, then each point within the matrix is N covered to within the CL-F distance $[L_{CM}, W_{CM}]$, wherein

13 L_{CM} is the length of a matrix cell and W_{CM} is the width of a matrix cell. A choice of covering markers so

14 that approximately the same number of covering markers are in each cell of a CL-F matrix has utility in

15 that approximately the same amount of effort is expended on each subregion (cell) of the CL-F region

16 defined by the matrix in a linkage study using the covering markers. If the centerpoints of a CL-F matrix

17 (a matrix centerpoint lattice) are each N covered by a group of covering markers to within a CL-F

18 distance [x,y], then each point in the matrix is N covered to within the CL-F distance [2x,2y]. A CL-F

19 matrix can be used as a device to help distinguish versions of the invention from prior art (to the extent

20 that there is prior art).

21 A requirement that the CL-F region that is N covered to within a certain CL-F distance comprise a

22 certain minimum area or segment-subrange with a certain minimum area is a special case of CL-F

23 regions that are N covered to within the certain CL-F distance. A requirement that the CL-F region that

24 is N covered to within a certain CL-F distance has a certain length or width is a special case of CL-F

25 regions that are N covered to within the certain CL-F distance. Each of these requirements is also a

26 device that can be used to help distinguish versions of the invention from prior art.

27 **A Note on the Equivalence of Working With Individual Alleles of Markers to Perform Two-**

28 **dimensional Linkage Studies and the CL-F approach using bi-allelic markers**

29 It is possible to conceptualize performing two-dimensional linkage studies wherein individual marker

30 alleles are used to cover a two-dimensional space, rather than individual bi-allelic markers. Any

31 individual marker allele is assigned a two-dimensional location consisting of the chromosomal location

32 of the marker and the allele frequency of the marker allele. Two-dimensional chromosomal location-

33 allele frequency spaces(or regions) are systematically covered by sets of covering alleles. Each

34 individual covering allele is tested for association with a genetic characteristic. Versions of inventions

35 using systematic chromosomal location-allele frequency(CL-AF) region coverings that are similar to

36 versions of the invention in this application are possible. Indeed these types of inventions have been

37 described in U.S.Provisional Patent Applications previously filed by the inventor.

38 However, such a conceptual framework and the resulting inventions are equivalent to the CL-F versions

39 approach used in this application. This is because any marker allele, A, that is used as a covering allele

40 can be made to be an allele equivalent of a bi-allelic marker equivalent(BME). So that a BME with allele

1 equivalents A and nonA is a bi-allelic marker with allele A. Therefore, any set of covering alleles that
2 systematically cover a two-dimensional CL-AF region is equivalent to a set of BMEs that systematically
3 cover an equivalent CL-F region. Testing each covering allele for association with a genetic
4 characteristic is exactly equivalent to testing each BME of a set of BMEs for evidence of linkage to a
5 gene using a statistical linkage test based on allelic association. Even testing for the presence or
6 absence of a covering allele in the chromosomal DNA of an individual is equivalent to genotyping the
7 individual at a BME. And determining a sample allele frequency for a covering allele, is equivalent to
8 determining the sample allele frequencies for a BME.
9
10
11 **Example 1 of Process #1 is used for identifying markers linked to a disease gene.**
12
13 Example 1 A process for identifying bi-allelic markers linked to a bi-allelic disease gene in human
14 beings, comprising:
15
16 a)choosing two or more bi-allelic covering markers so that a CL-F region is N covered to within a CL-F
17 distance [250,000 bp, 0.1] or the equivalent thereof by the covering markers, wherein N is an integer
18 number greater than or equal to 2 ;
19
20 b)choosing the same statistical linkage test based on allelic association for each covering marker;
21
22 c)choosing the same sample of individual human beings for each covering marker;
23
24 d)obtaining genotype data at each covering marker for each individual in the sample and obtaining
25 phenotype status data for the disease for each individual in the sample ;
26
27 e)calculating evidence for linkage between each covering marker and the gene using the test chosen in
28 b) and the genotype data at each covering marker and the using the phenotype status data for the
29 disease for each individual in the sample ; and
30
31 f)identifying those covering markers as linked to the gene which show evidence for linkage based on
32 the calculations of e.

Apparatus Versions

34
35 <u>**General descriptions of individual apparatus versions are given below.**</u>
36
37 **Apparatus #1, an apparatus to practice process #1.**
38
39

1   **Apparatus #1**, An apparatus for identifying bi-allelic markers linked to a bi-allelic genetic characteristic

2   gene in a species of creatures, comprising :

3

4   a)means for choosing two or more bi-allelic covering markers so that a CL-F region is systematically

5   covered by the two or more covering markers;

6

7   b)means for choosing a statistical linkage test based on allelic association for each covering marker;

8

9   c)means for choosing a sample of individuals for each covering marker ;

10

11   d)means for obtaining genotype data/sample allele frequency data for each covering marker and the

12   sample chosen for each covering marker, and for obtaining phenotype status data for the genetic

13   characteristic for each individual in the sample chosen for each covering marker;

14

15

16   e)means for calculating evidence for linkage between each covering marker and the gene using the

17   statistical linkage test based on allelic association chosen for each covering marker and the genotype

18   data/sample allele frequency data for each covering marker and using the phenotype status data for the

19   genetic characteristic for each individual in the sample chosen for each covering marker obtained in d);

20   and

21

22

23   f)means for identifying those covering markers as linked to the gene which show evidence for linkage

24   based on the calculations by means e).

25

26   **More detailed description of Apparatus #1:** Apparatus #1 is an apparatus to practice

27   process #1. More details of the description of apparatus #1 are found under the description of Process

28   #1 above. Any one of the means labeled a), b), c), d), e) or f) of apparatus #1 includes any means for

29   automating or partially automating a step or step-like part as a), b), c), d), e) or f) respectively of

30   process #1. An example of any one of the means in this paragraph labeled a), b), c), d), e), or f) is

31   means comprising an appropriately programmed, suitable computer, the computer being supplied with

32   proper data and instructions.

33

34   The means labeled d) of apparatus #1 for obtaining genotype data/ sample allele frequency data for

35   each covering marker for the sample chosen for each covering marker includes any automated or

36   partially automated means to obtain genotype data/ sample allele frequency data. An example of

37   means to obtain genotype data/ sample allele frequency data is means using mass spectrometry.[1]

38   Means to obtain genotype data/ sample allele frequency data that is automated or partially automated

39   includes means comprising Oligonucleotide Technology described below.

40   **Apparatus #1A, an apparatus to practice process #1A.**

1

2 **Apparatus#1A** : An apparatus for localizing a bi-allelic genetic characteristic gene in a species

3 of creatures to a chromosomal location-least common allele frequency (CL-F) region,

4 comprising the means a), b), c), d) and e) of Apparatus #1 and further comprising the means of:

5 f)means for localizing the gene to the approximate chromosomal location-least common allele

6 frequency region (CL-F)of one or more markers that show evidence for linkage based on the

7 calculations of means e).

8 An example of means f) is means comprising an appropriately programmed, suitable computer, the

9 computer being supplied with proper data and instructions. Further details of this apparatus which

10 practices process #1A are under process #1 and process #1A and **Software**(above).

11

12 ## Genotype data/Sample allele frequency data apparatus

13 An apparatus to obtain genotype data/sample allele frequency data similar to the data of the d) of

14 process #1 has great utility in that it is used to provide genotype data /sample allele frequency data for

15 the more powerful two-dimensional linkage studies introduced in this application.

16 **ApparatusGd/Safd#1: Genotype data/Sample allele frequency data apparatus: An apparatus for**

17 **obtaining genotype data/sample allele frequency data for each bi-allelic marker of a group of**

18 **two or more bi-allelic covering markers in the chromosomal DNA of one or more individuals of a**

19 **sample, comprising:**

20 **a) means for determining information on the presence or absence of each allele of each bi-allelic**

21 **marker of a group of two or more bi-allelic covering markers in the chromosomal DNA of one or**

22 **more individuals of the sample, a CL-F region being systematically covered by the two or more**

23 **bi-allelic covering markers; and**

24 **b) means for transforming the information of a) into genotype data/sample allele frequency data**

25 **for each marker of the group.**

26 The CL-F region and covering markers are for a species and the one or more individuals are members

27 of the species. Means for determining information on the presence or absence of each allele of each bi-

28 allelic marker of the group in chromosomal DNA includes any means of determination. Means for

29 determining information on the presence or absence of each allele of each bi-allelic marker of the group

30 in chromosomal DNA includes means comprising oligonucleotide technology by using a set of

31 oligonucleotides that is complementary to the group as discussed below. Information on the presence

32 or absence of each allele in the chromosomal DNA is obtained using a DNA specimen from each of one

33 or more individuals of the sample or by using one or more DNA pools of DNA specimens from two or

34 more individuals of the sample. Any apparatus that obtains genotype data or sample allele frequency

35 data (similar to the data of the d) of process #1) by determining the presence or absence of each allele

36 of each bi-allelic marker of the group in the chromosomal DNA of one or more individuals is an example

37 of this version of the invention. Versions of this apparatus also obtain a combination of genotype data

1 and sample allele frequency data similar to the data of the d) of process #1. The details of b) will be

2 clear to those of ordinary skill in the art.

3

4 Each bi-allelic covering marker is a true bi-allelic or BME. Determining the presence or absence of each

5 allele of each bi-allelic marker in the group includes determining the presence or absence of each allele

6 equivalent of each bi-allelic marker equivalent(BME) in the group. Any method of systematically

7 covering the CL-F region is acceptable. In this application, the systematic covering of a CL-F region in

8 versions of the invention is described mathematically as the covering of a CL-F region, wherein the CL-

9 F region is N covered to within a CL-F distance δ by two or more bi-allelic covering markers. For further

10 details regarding this, see Detailed Description of the Systematic Covering of a CL-F Region Used In

11 Versions of the Invention above.

12 **An example of ApparatusGd/Safd#1Genotype data/Sample allele frequency data apparatus, a**

13 **sample allele frequency apparatus:**

14 Example 1 of ApparatusGd/Safd#1:An apparatus for obtaining genotype data/sample allele frequency

15 data for each bi-allelic marker of a group of two or more bi-allelic covering markers in the chromosomal

16 DNA of one or more individuals of a sample, wherein the genotype data/sample allele frequency data is

17 sample allele frequency data, comprising:

18 a) means for determining information on the presence or absence of each allele of each bi-allelic

19 marker of a group of two or more bi-allelic covering markers in the chromosomal DNA from one or more

20 individuals of the sample, a CL-F region being N covered to within the CL-F distance [1.0 cM, 0.15] by

21 the two or more bi-allelic covering markers, wherein N is an integer number greater than or equal to 1;

22 and

23 b) means for transforming the information of a) into sample allele frequency data for each marker of the

24 group.

25 Example 2 of ApparatusGd/Safd#1:An apparatus for obtaining genotype data/sample allele frequency

26 data for each bi-allelic marker of a group of two or more bi-allelic covering markers in the chromosomal

27 DNA of an individual, wherein the genotype data/sample allele frequency data is genotype data,

28 comprising:

29 a) means for determining information on the presence or absence of each allele of each bi-allelic

30 marker of a group of two or more bi-allelic covering markers in the chromosomal DNA from an

31 individual, a CL-F region being N covered to within the CL-F distance [12cM, 0.25] or the equivalent

32 thereof by the two or more bi-allelic covering markers, wherein N is an integer number greater than or
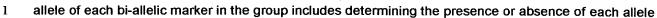
33 equal to 1; and

34 b) means for transforming the information of a) into genotype data for each marker of the group.

1 (It should be noted that the following genotype apparatus is equivalent to Example 2 of

2 ApparatusGd/Safd#1: Genotype Apparatus: An apparatus for genotyping an individual, comprising:

3 a)means to genotype an individual at two or more bi-allelic covering markers, a CL-F region being N

4 covered to within the CL-F distance [12cM, 0.25] or the equivalent thereof by the two or more bi-allelic

5 covering markers, wherein N is an integer number greater than or equal to 1.)

6

7 ## Genotype data/Sample allele frequency data process

8 A process to obtain genotype data/sample allele frequency data similar to the data of the d) of process

9 #1 has great utility in that it is used to provide genotype data /sample allele frequency data for the more

10 powerful two-dimensional linkage studies introduced in this application.

11 **Description of the Genotype data/Sample allele frequency data process.**

12 **ProcessGd/Safd#1: Genotype data/Sample allele frequency data process: A process for**

13 **obtaining genotype data/sample allele frequency data for each bi-allelic marker of a group of**

14 **two or more bi-allelic covering markers in the chromosomal DNA of one or more individuals of a**

15 **sample, comprising:**

16 **a) determining information on the presence or absence of each allele of each bi-allelic marker of**

17 **a group of two or more bi-allelic covering markers in the chromosomal DNA of one or more**

18 **individuals of the sample, a CL-F region being systematically covered by the two or more bi-**

19 **allelic covering markers; and**

20 **b) transforming the information of a) into genotype data/sample allele frequency data for each**

21 **marker of the group.**

22 The CL-F region and covering markers are for a species and the one or more individuals are members

23 of the species. Determining information on the presence or absence of each allele of each bi-allelic

24 marker of the group in chromosomal DNA includes any method of determination. Determining

25 information on the presence or absence of each allele of each bi-allelic marker of the group in

26 chromosomal DNA includes methods comprising oligonucleotide technology by using a set of

27 oligonucleotides that is complementary to the group as discussed below. Information on the presence

28 or absence of each allele in the chromosomal DNA is obtained using a DNA specimen from each of one

29 or more individuals of the sample or by using one or more DNA pools of DNA specimens from two or

30 more individuals of the sample. Any process that obtains genotype data or sample allele frequency data

31 (similar to the data of the d) of process #1) by determining the presence or absence of each allele of

32 each bi-allelic marker of the group in the chromosomal DNA of one or more individuals is an example of

33 this version of the invention. Versions of this process also obtain a combination of genotype data and

34 sample allele frequency data similar to the data of the d) of process #1. The details of b) will be clear to

35 those of ordinary skill in the art.

36

37 Each bi-allelic covering marker is a true bi-allelic or BME. Determining the presence or absence of each

1 allele of each bi-allelic marker in the group includes determining the presence or absence of each allele

2 equivalent of each bi-allelic marker equivalent(BME) in the group. Any method of systematically

3 covering the CL-F region is acceptable. In this application, the systematic covering of a CL-F region in

4 versions of the invention is described mathematically as the covering of a CL-F region, wherein the CL-

5 F region is N covered to within a CL-F distance δ by two or more bi-allelic covering markers. For further

6 details regarding this, see Detailed Description of the Systematic Covering of a CL-F Region Used In

7 Versions of the Invention above.

8 **An example of ProcessGd/Safd#1Genotype data/Sample allele frequency data process, a**

9 **genotype data process:**

10 Example 1 of ProcessGd/Safd#1:A process for obtaining genotype data/sample allele frequency data

11 for each bi-allelic marker of a group of two or more bi-allelic covering markers in the chromosomal DNA

12 of an individual, wherein the genotype data/sample allele frequency data is genotype data, comprising:

13 a) determining information on the presence or absence of each allele of each bi-allelic

14 marker of a group of two or more bi-allelic covering markers in the chromosomal DNA from an

15 individual, a CL-F region being N covered to within the CL-F distance [12cM, 0.25] or the equivalent

16 thereof by the two or more bi-allelic covering markers; wherein N is an integer number greater than or

17 equal to 1; and

18 b) transforming the information of a) into genotype data for each marker of the group.

19 (It should be noted that the following genotype process is equivalent to Example 1 of

20 ProcessGd/Safd#1: Genotype Process: A process for genotyping an individual, comprising:

21 a) genotyping an individual at two or more bi-allelic covering markers, a CL-F region being N covered to

22 within the CL-F distance [12cM, 0.25] or the equivalent thereof by the two or more bi-allelic covering

23 markers, wherein N is an integer number greater than or equal to 1.)

24

25 **Oligonucleotide technology**

26 Each version of oligonucleotide technology is a means to sense the presence or absence of each of

27 one or more true alleles of a group of true alleles in chromosomal DNA from one or more individuals by

28 means of a hybridization reaction with an oligonucleotide that is complementary to each of the one or

29 more true alleles (see definitions section). Thus versions of oligonucleotide technology are a means of

30 genotyping one or more individuals. And, versions of oligonucleotide technology are a means of

31 obtaining sample allele frequency data for one or more marker alleles for a sample of individuals using

32 pooled DNA from the individuals in the sample.

33 In Some Versions of Oligonucleotide Technology for Genotyping or Obtaining Sample Allele Frequency

34 Data, a Physico-chemical Signal is Generated when an Allele in Chromosomal DNA and a

35 Complementary Oligonucleotide Hybridize

1 Some versions of oligonucleotide technology for genotyping or for obtaining sample allele frequency

2 data use a sensor which includes one or more oligonucleotides which are complementary to an allele.

3 When the sensor is exposed to chromosomal DNA from an individual who carries the allele, the

4 oligonucleotides which are complementary to the allele hybridize with chromosomal DNA specimens of

5 the allele. The hybridization generates a physico-chemical signal which indicates the presence of the

6 allele in the chromosomal DNA of the individual. The lack of the physico-chemical signal indicates no

7 (or negligible)hybridization and that the allele is not present in the chromosomal DNA of an individual.

8 Examples of oligonucleotide technology for genotyping, obtaining sample allele frequency data or

9 genotype data/sample allele frequency data

10 Companies like Affymetrix are using high density arrays of oligonucleotides attached to silicon chips or

11 glass slides to genotype DNA from one individual at thousands of bi-allelic markers.[8] In some of these

12 versions of oligonucleotide technology, the strength of hybridization of oligonucleotides that differ at

13 only one base to DNA containing an SNPare compared to determine genotype.[9] Another version of

14 oligonucleotide technology uses oligonucleotides as PCR (Polymerase Chain Reaction) primers to

15 obtain genotype data.[10]Other examples of oligonucleotide technology and it's uses to obtain genetic

16 information are included in the articles cited in the endnotes.[11] Versions of oligonucleotide technology

17 obtain sample allele frequency data from pooled DNA or genotype data using oligonucleotides as PCR

18 primers to obtain amplified reaction products that are detected by mass spectrometry. Another example

19 of oligonucleotide technology is padlock probes.[12]

20 Other examples of oligonucleotide technology are minisequencing on DNA arrays, dynamic allele-

21 specific hydridization, microplate array diagonal gel electrophoresis, pyrosequencing, oligonucleotide-

22 specific ligation, the TaqMan system and immobilized padlock probes as presented at the First

23 International Meeting on Single Nucleotide Polymorphism and Complex Genome Analysis.[13]

24 Sets of Oligonucleotides for Genotyping at Bi-allelic Markers or Obtaining Sample Allele Frequency

25 Data

26 A *set of oligonucleotides that is complementary(see definitions) to a group of one or more bi-allelic*

27 *markers has utility to determine genotype data at each of the markers in the group*, including groups

28 with BMEs and approximately bi-allelic markers.

29 Similarly, a set of oligonucleotides that is complementary to a group of bi-allelic markers has utility to

30 obtain sample allele frequency data for each allele of each marker in the group.

31 *In both cases, obtaining genotype data or sample allele frequency data, the same principle is*

32 *used: a set of oligonucleotides that is complementary to a group of bi-allelic markers has utility*

33 *to determine the presence or absence of each allele of each marker in the group in*

34 *chromosomal DNA.*

35 Using sets of oligonucleotides to obtain Genotype Data/Sample Allele Frequency Data for each

36 marker of a group of bi-allelic markers, wherein the group of markers systematically cover a CL-

37 F region

38 Genotype data/sample allele frequency data for each marker of a group of bi-allelic markers, wherein

39 the group of bi-allelic markers systemically cover a CL-F region has great utility for use in the more

40 powerful two-dimensional linkage studies introduced in this application. As described above under

1 Oligonucleotide Technology, some sets of oligonucleotides have utility to determine genotype data at
2 each bi-allelic marker of a group of one or more bi-allelic markers. Similarly, some sets of
3 oligonucleotides have utility to obtain sample allele frequency data for each bi-allelic marker of a group
4 of one or more bi-allelic markers. Therefore, the use of one or more copies of a set of oligonucleotides
5 to obtain genotype data or sample allele frequency data for each bi-allelic marker of a group of one or
6 more bi-allelic covering markers, wherein the group of bi-allelic covering markers systematically cover a
7 CL-F region has great utility.
8 A word to avoid confusion in terminology: in this application, a set of markers for use in genotyping is
9 referred to as a set of oligonucleotides.
10 A set of oligonucleotides consisting of one or both strands of each allele of a group of one or more
11 markers is a set of oligonucleotides that is complementary to the group of markers. (see definitions
12 section) Such a set of oligonucleotides is in effect the group of markers themselves; and such a set of
13 oligonucleotides has utility to determine genotype data at each marker in the group. So a group of
14 markers (or set of markers) for use in obtaining genotype data or sample allele frequency data for each
15 of the markers in the group is included in the descriptive phrase: "a set of oligonucleotides".
16 **Description of Use set#1 D:**
17 **Use set#1 D The use of one or more copies of a set of oligonucleotides to determine genotype**
18 **data/sample allele frequency data for each bi-allelic marker of a group of two or more bi-allelic**
19 **covering markers for one or more individuals, wherein the group of covering markers**
20 **systematically cover a CL-F region.**
21 The CL-F region and covering markers are for a species and the one or more individuals are members
22 of the species. An example of a set of oligonucleotides with utility to be used to determine genotype
23 data/sample allele frequency data for each bi-allelic marker of a group of two or more bi-allelic covering
24 markers is a set of oligonucleotides that is complementary to the group of markers. A set that is
25 complementary to the group of markers is used to detect the presence or absence of each the alleles of
26 the covering markers by means of a hybridization reaction as discussed under oligonucleotide
27 technology. Thus a set that is complementary to the group of markers is used to determine genotype
28 data/sample allele frequency data for each covering marker.
29 The use of one or more copies of a set of oligonucleotides to obtain genotype data or sample allele
30 frequency data for each bi-allelic marker of a group of one or more bi-allelic covering markers, wherein
31 the group of bi-allelic covering markers systematically cover a CL-F region are both examples of this
32 version of the invention(Use Set#1D).
33 In this application, the systematic covering of a CL-F region in versions of the invention is described
34 mathematically as the covering of a CL-F region, wherein the CL-F region is N covered to within a CL-F
35 distance δ by two or more bi-allelic covering markers. For further details regarding this, see Detailed
36 Description of the Systematic Covering of a CL-F Region Used In Versions of the Invention above.
37 **Example 1S of Use set#1D:** The use in genotyping one or more individuals, of one or more
38 copies of a set of oligonucleotides, the set of oligonucleotides being complementary to a group of two or
39 more bi-allelic covering markers, a CL-F region being N covered by the covering markers to within a

●   ●

1 CL-F distance of about [250,000 bp, 0.1] or the equivalent thereof, wherein N is an integer greater than

2 or equal to two.

3 **Composition of matter: Description of Comp set#1D:**

4

5 **Comp set#1D: One or more copies of a set of oligonucleotides, the set of oligonucleotides being**

6 **complementary to a group of two or more bi-allelic covering markers, wherein the group of**

7 **covering markers systematically cover a CL-F region.**

8 A set of oligonucleotides that is complementary to a group of two or more bi-allelic covering markers,

9 wherein the group of covering markers systematically cover a CL-F region has great utility for use in the

10 two-dimensional linkage study techniques introduced in this application. Such a set has utility in being

11 used to genotype individuals or obtain sample allele frequency data or genotype data/sample allele

12 frequency data as described above under Use set#1D. In this application, the systematic covering of a

13 CL-F region in versions of the invention is described mathematically as the covering of a CL-F region,

14 wherein the CL-F region is N covered to within a CL-F distance δ by two or more bi-allelic covering

15 markers. For further details regarding this, see <u>Detailed Description of the Systematic Covering of a CL-</u>

16 <u>F Region Used In Versions of the Invention</u> above.

17

18 ## Example 1Comp of Comp set#1D:

19 **Example 1Comp:** One or more copies of a set of oligonucleotides, the set of oligonucleotides being

20 complementary to a group of two or more bi-allelic covering markers, a CL-F region being N covered by

21 the covering markers to within a CL-F distance of about [1cM, 0.2] or the equivalent thereof, wherein N

22 is an integer greater than or equal to one.

23 ## Redundancy of Covering Markers

24 Some versions of the invention make use of N coverings of CL-F regions by covering markers which

25 limit (possibly to zero) the number of pairs of covering markers which are redundant within CL-F

26 distance D, D = [D_{CL}, D_F], wherein D is less than or equal to about δ, a CL-F covering distance. This

27 limits the number covering markers which are separated by a CL-F distance of less than or equal to D(if

28 the markers were placed on a CL-F map) which *will be in extreme positive disequilibrium with each*

29 *other.* This limitation is done by requiring that less than or equal to R pairs of covering markers are

30 redundant within distance D. Wherein R is an integer greater than or equal to 0 and less than or equal

31 to about N(N-1)/2. When R is chosen to be zero, no pair of covering markers is redundant within

32 distance D.

33 A preferable condition is that each bi-allelic covering marker within each small CL-F region (a small

34 segment-subrange of length about $δ_{CL}$ and width about $δ_F$ the distance components of the covering

35 distance δ) provides much new (i.e. non-redundant) information about linkage and association to any

36 nearby bi-allelic gene. Under these conditions, testing each bi-allelic covering marker in each small CL-

37 F region increases the likelihood of detecting linkage to a gene.

38 Limiting (including to zero) pairs of covering markers which are redundant within CL-F distance D(which

39 is less than or equal to a covering distance δ) approaches and achieves this preferable condition. This

1 limitation is not crucial to the functioning of a version of the invention, however, it has the advantage of

2 reducing excess effort and increasing efficiency.

Polymorphism CL-F, Linkage Disequilibrium and Association Data Display

5 Polymorphism CL-F display apparatus display the chromosomal location, least common allele

6 frequency and identity of each polymorphism of one or more polymorphisms (markers or genes or both)

7 of one or more populations of one or more species on one or more two-dimensional graphs, each graph

8 is similar to an x-y plot. In addition, versions of this invention display known or estimated Linkage

9 Disequilibrium (LD) data between two or more polymorphisms for one or more populations. Other

10 versions of the invention display known or estimated association data between two or more alleles of

11 two or more polymorphisms for one or more populations. The apparatus has utility including aiding in

12 decisions regarding linkage studies and the interpretation of linkage study data.

13 The apparatus comprise means to display the chromosomal location, least common allele frequency

14 and identity of each polymorphism of one or more polymorphisms (markers or genes or both) of one or

15 more populations of one or more species on one or more two-dimensional graphs, each graph is similar

16 to an x-y plot. Some versions of this apparatus display LD data or association data (or both) between

17 two or more polymorphisms or two or more alleles of polymorphisms respectively for one or more

18 populations of one or more species on one or more two-dimensional graphs, wherein each graph is

19 similar to an x-y plot and indicates the CL-F location and identity of each polymorphism. In some

20 versions, human interaction from a human operator with a mouse or similar device causes the LD data

21 or association data (or both) between each of one or more pairs of polymorphisms to be displayed.

22 Each graph has two axes, one axis, the frequency axis, represents least common allele frequency and

23 the alternate(or other) axis, the chromosomal location axis, represents chromosomal location. Each

24 frequency axis of each graph is in units of population frequency. Each chromosomal location axis of

25 each graph is in units of chromosomal location such as centimorgans, base pairs or the equivalent

26 thereof.

27 The frequency axis of each graph spans the entire range 0 to 0.5 or a subrange of the range 0 to 0.5.

28 The chromosomal location axis of each graph spans the chromosomal locations on one or more

29 segments of one or more chromosomes of a species, each of the one or more segments is a size from

30 the equivalent of a base pair in length to the length of an entire chromosome (or the equivalent thereof).

31 Each point on each graph is directly opposite a value on the frequency axis of each graph. The value

32 on the frequency axis directly opposite each point on each graph is the frequency coordinate of each

33 point on each graph. Each point on each graph is directly opposite a value on the chromosomal location

34 axis of each graph. The value on the chromosomal location axis directly opposite each point on each

35 graph is the chromosomal location coordinate of each point on each graph.

36 Each graph displays the chromosomal location and least common allele frequency of each

37 polymorphism of one or more polymorphisms. Each polymorphism displayed on each graph is assigned

38 a graph location on each graph.

39 The graph location of each polymorphism displayed on each graph is typical of the use of x-y plots. The

40 graph location assigned to each polymorphism on each graph is a point. The chromosomal location

41 coordinate of the point assigned as the graph location to any one polymorphism is equal (or

1 approximately equal) to the chromosomal location of the polymorphism. And the frequency coordinate

2 of the point assigned as the graph location to any one polymorphism is equal (or approximately equal)

3 to the least common allele frequency of the polymorphism.

4 The apparatus comprise means for displaying one or more two-dimensional graphs. Each graph

5 comprises, the identity and graph location of one or more polymorphisms assigned a location on each

6 graph. And the apparatus comprise means for displaying one or more graphs wherein the viewer

7 chooses the species, population, polymorphisms, span of the frequency axis and span of the

8 chromosomal location axis of the one or more graphs ; in versions of the apparatus, the means of this

9 sentence comprises a computer.

10 The apparatus comprise means for storing and updating data on the chromosomal location and least

11 common allele frequency of one or more polymorphisms of one or more populations of one or more

12 species and means for storing chromosomal location and least common allele frequency data on newly

13 discovered polymorphisms.

14 Versions of the apparatus comprise means for printing each of the one or more graphs.

15 **Theory of Operation / Set/Subset Example**

16 Systematically Varying Both Marker Chromosomal Location and Marker Allele Frequency of Markers in

17 Linkage Studies

18 The inventor's calculations and observations have demonstrated the increased power of the TDT in

19 more common, less optimal situations when a bi-allelic marker and bi-allelic gene have (1) similar but

20 not identical allele frequencies and (2) the marker and gene are in some degree of linkage

21 disequilibrium. Thus, for a typical linkage study using bi- allelic markers and an association based

22 linkage test, *to increase the likelihood of both criteria (1) and (2) occurring for one or more*

23 *markers, so as to increase the power of an association based linkage test in a linkage study, the*

24 *bi-allelic markers used in the study are chosen so that the least common allele frequencies of*

25 *the markers vary systematically over a range or subrange of least common allele frequency*

26 *AND the chromosomal location of the markers vary systematically over one or more*

27 *chromosomes or chromosomal regions. And the bi-allelic markers are chosen so that the*

28 *markers' chromosomal locations and least common allele frequencies vary systematically in an*

29 *essentially independent manner.*

30 (In theTheory of Operation/ Set/Subset Example Section the traditional symbol used in scientific papers

31 for the disequilibrium coefficient, δ, is used. This should not be confused with the symbol δ used for the

32 covering distance in the remainder of the application. The symbol d is used for the disequilibrium

33 coefficient in the sections of the application other than the Theory of Operation/Set/Subset Example

34 Section.)

35 The theory of operation is based on the mathematical observation that the TDT and other association-

36 based tests for linkage are increased in power as the frequencies of the disease-causing allele of a bi-

37 allelic gene and the positively associated allele of a linked bi-allelic marker become similar in

38 magnitude. The inventor made this observation as a result of deriving the equation shown below for P_t

39 (this is Equation 2 in the unpublished manuscript submitted for publication in December 1996 and in

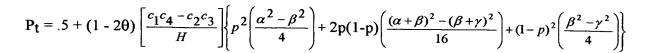40 published paper by RE McGinnis in the Annals of Human Genetics vol 62, pp. 159-179, 1998).

1

$$P_t = .5 + (1 - 2\theta) \left[\frac{c_1 c_4 - c_2 c_3}{H} \right] \left\{ p^2 \left(\frac{\alpha^2 - \beta^2}{4} \right) + 2p(1-p) \left(\frac{(\alpha + \beta)^2 - (\beta + \gamma)^2}{16} \right) + (1 - p)^2 \left(\frac{\beta^2 - \gamma^2}{4} \right) \right\}$$

2   Equation 2

3

4   P_t may be regarded as the size of the "signal" which is given by the TDT to indicate that a tested

5   marker is linked to a disease-causing gene. The more P_t is elevated above 0.5 (baseline), the greater is

6   the evidence for linkage or "power" provided by the association-based linkage test known as the TDT.

7   Table 2 in the unpublished manuscript filed with previous US Provisional Patent

8   Applications(see below) illustrates how signal strength increases substantially as the frequencies of

9   disease-causing allele and positively associated marker allele become similar in magnitude. As noted

10   on pages 24 and 25 of the unpublished manuscript(see below), Table 2 assumes that the frequency (p)

11   of the disease-causing allele is fixed at p=.1 while the frequency (m) of the positively associated marker

12   allele varies (m=.5, .3, .2, .1, .05). Note that when the level of disequilibrium (or association) between

13   the bi-allelic marker and bi-allelic disease gene is fixed (in this case either $\delta = \delta_{max}$ or $\delta = \frac{1}{2} \delta_{max}$), the

14   signal strength of P_t progressively increases as m decreases from m=.5 to m=.1 (the same frequency

15   as the disease allele, i.e., p=.1). For example, in the section of Table 2 for r=5, note that when $\delta = \frac{1}{2}$

16   δ_{max}, P_t is .548 at m=.5 and then steadily increases to .572 (m=3), .597 (m=.2), .648 (m=.1) and then

17   starts to decrease again as m departs from m=p=.1 (i.e. P_t =.636 at m=.05). As noted on pages 24-25

18   (below)of the unpublished manuscript, the TDT chi-square statistic (assuming a sample size of 200

19   families) is such that the signal strength at m=.5 (P_t =.548) does not produce a statistically significant

20   evidence for linkage (p-value > 0.05) while the doubling of signal strength at m=.2 (P_t =.597) produces

21   very strong statistical evidence for linkage by the TDT (p-value< 0.005). This sort of substantial

22   increase in power is also true of other association-based linkage tests as the frequencies of the

23   disease-causing allele and associated marker allele become more similar in magnitude.

24

1 Table 2(Footnotes for Table 2 are on next page)

2 Effect of penetrance ratio (r), disequilibrium (δ) and marker heterozygosity (m) on magnitude

3 of P_t and P_s

		Magnitude of P_t			Magnitude of P_s		
		δ_{max}[a]	$\frac{1}{2}\delta_{max}$[b]	$\delta=0$	δ_{max}[a]	$\frac{1}{2}\delta_{max}$[b]	$\delta=0$
r=2	m=.5	.526	.513	.500	.505	.505	.504
	m=.3	.541	.521	.500	.508	.506	.504
	m=.2	.558	.531	.500	.511	.508	.504
	m=.1	.595	.555	.500	.518	.512	.504
	m=.05	.589	.552	.500	.517	.511	.504
r=5	m=.5	.596	.548	.500	.543	.540	.539
	m=.3	.633	.572	.500	.561	.548	.539
	m=.2	.666	.597	.500	.575	.556	.539
	m=.1	.719	.648	.500	.600	.573	.539
	m=.05	.696	.636	.500	.589	.571	.539
r=10	m=.5	.656	.577	.500	.595	.587	.584
	m=.3	.702	.612	.500	.623	.600	.584
	m=.2	.736	.644	.500	.644	.612	.584
	m=.1	.785	.703	.500	.673	.635	.584
	m=.05	.750	.684	.500	.652	.628	.584
r=∞	m=.5	.740	.617	.500	.700	.680	.673
	m=.3	.791	.663	.500	.743	.700	.673
	m=.2	.826	.703	.500	.772	.716	.673
	m=.1	.870	.770	.500	.807	.744	.673
	m=.05	.816	.741	.500	.763	.730	.673

1 Footnotes for Table 2

2 [a,b]Value of δ that is maximal (δ_{max}) and half-maximal ($\frac{1}{2}\delta_{max}$), as determined by the

3 heterozygosity of the marker (m) and disease locus (p=.1)

4

5 <u>Importance of disequilibrium and marker heterozygosity (i.e. marker allele frequency) in</u>

6 <u>detecting linkage</u>

7   When the heterozygosity (i.e. allele frequencies) of a bi-allelic marker and bi-allelic

8 disease locus are fixed, $(P_S - .5)$ and $|P_t - .5|$ are both maximized at the most positive or most

9 negative possible value of δ (δ_{max}, δ_{min}), as demonstrated in the published paper. This

10 maximization of χ^2_{asp} and χ^2_{tdt} is intimately connected to M_S and M_t (defined in equations 1

11 and 2) since: (a) these are the only two factors in P_S and P_t that are influenced by δ and (b) M_S

12 and $|M_t|$ are maximal and equal to each other when δ is extreme (δ_{max} or δ_{min}). Furthermore, as

13 explained in the published paper, M_S is a measure of the proportion of informative (A/B)

14 parents who are also informative (D/d) at the disease locus. Therefore, maximizing M_S (and,

15 by implication, $|M_t|$) is equivalent to *minimizing* the proportion of A/B parents who are

16 homozygous (D/D or d/d) at the disease locus. Such homozygous D/D or d/d parents

17 contribute evidence *against* linkage since they transmit marker alleles A and B to affected

18 offspring with equal probability; thus, minimizing their proportion among A/B parents being

19 tested for linkage has the effect of maximizing χ^2_{asp} and χ^2_{tdt}.

20   Nevertheless, when bi-allelic markers have a specific (i.e. fixed) heterozygosity

21 different from that of a bi-allelic disease locus, some A/B parents must be homozygous at the

22 disease locus, even when δ is extreme. But if marker heterozygosity is variable, the proportion

23 of A/B parents who are D/D or d/d approaches *zero* as marker heterozygosity approaches that

24 of the disease locus and as δ approaches δ_{max} or δ_{min}. Consequently, the most extreme values

25 of P_t and P_S, and highest values of χ^2_{tdt} and χ^2_{asp} are found when marker and disease locus

26 have equal heterozygosity and $\delta=\delta_{max}$ or $\delta=\delta_{min}$.

27

28 <u>Example illustrating the importance of marker heterozygosity (i.e. allele frequency)</u>

1   To illustrate the importance of marker heterozygosity and disequilibrium, Table 2

2 shows P_t and P_s values when the frequency (p) of disease allele D is constant at 0.1, but the

3 frequency (m) of marker allele A varies between m=.5 (maximum marker heterozygosity) and

4 m=.1 (equal heterozygosity at marker and disease loci). The table assumes mode of

5 inheritance is additive, and separate sections of the table show the results when the penetrance

6 ratio (r) is 2, 5, 10 or ∞. For each value of r, an individual line in the table represents constant

7 marker heterozygosity (m=.5, .3, .2, or .1) and from left-to-right on each line, one sees P_t and

8 P_s values when $\delta=\delta_{max}$, $\delta=\frac{1}{2}\delta_{max}$, and $\delta=0$, the value of δ_{max} being determined by the particular

9 values of m and p [$\delta_{max}=p(1-m)$]. As noted in Appendix I of the published paper, when p<m

10 and p<(1-m), as in this example, the most extreme values of P_t and P_s must occur at $\delta=\delta_{max}$.

11 This can be seen in each line of the table by the steady increase in both P_t and P_s as one moves

12 from $\delta=0$ to $\delta=\delta_{max}$, with every line also showing $P_t > P_s$ at $\delta=\delta_{max}$ and most lines showing $P_t >$

13 P_s at $\delta=\frac{1}{2}\delta_{max}$.

14   Most remarkable, however, are the sizeable increases in P_s and even greater increases

15 in P_t as marker heterozygosity drops toward the heterozygosity of the disease locus (m→.1). A

16 typical example is at r=5 and $\delta=\frac{1}{2}\delta_{max}$ where the table shows $P_t=.548$ at maximum marker

17 heterozygosity (m=.5) and $P_t=.597$ or .648 for m=.2 or .1, respectively. The impact of such an

18 increase in P_t can be understood by calculating χ^2_{tdt} for $P_t=.548$ (m=.5) and for $P_t=.597$ (m=.2)

19 assuming a data set of 200 families each with two affected sibs. Based on the expression for $\frac{H}{F}$

20 , I calculate the proportion of A/B parents to be .50 and .39 when m=.5 and .2, respectively.

21 So for m=.5, there would be .5 x 400 x 2 = 400 informative transmissions to affected offspring

22 with transmissions of allele A totaling .548 x 400 = 219, thus implying $\chi^2_{tdt}=\frac{38^2}{400} = 3.61$,

23 p<0.1. For m=.2, there would be .39 x 400 x 2 = 312 informative transmissions of which .597 x

24 312 = 186 would be transmissions of allele A yielding $\chi^2_{tdt}=\frac{60^2}{312}$ =11.54, p<0.005.

25   This example is typical, and highlights perhaps the most important finding of this

26 paper; namely the importance of using bi-allelic markers with heterozygosity similar to that of

27 a bi-allelic disease locus. Indeed, since a majority of susceptibility loci may be bi-allelic, the

44

1 judicious use of bi-allelic markers of both high, medium, and low heterozygosity may be

2 crucial in order to initially detect and replicate linkages to loci conferring modest disease risk.

3

4 **Set/Subset Example:**

5 Method for locating disease causing polymorphism using biallelic linkage
6 analysis

7

8

9 Objective : To test, by association-based linkage analysis (e.g., by TDT), whether a

10 disease-causing polymorphism is located on a particular chromosome (e.g., human

11 chromosome 4) or within a particular subregion of that chromosome.

12

13

14 PART 1 - Steps in conducting the association-based linkage test

15

16 Step 1

17 To conduct the test, first divide the chromosome or subregion of interest into segments

18 that are short enough that polymorphisms within each segment are likely to be in linkage

19 disequilibrium with each other. The division of a chromosome or subregion of interest into

20 "segments" is conceptual (*not* physical) and is based on chromosomal maps such as those

21 provided by the Whitehead Institute or Marshfield Foundation for Biomedical Research.

22 Although disequilibrium has been observed in Finnish populations between polymorphisms

23 that are 7 to 10 centimorgans (cM) apart, the chromosomal segments for searching for disease-

24 causing polymorphisms in more genetically heterogeneous populations should be less than 1

25 cM long (e.g., 250,000 base pairs long). These chromosomal segments might or might not

26 overlap each other (i.e., share some of their length in common); but the set of chromosomal

27 segments should completely cover the entire chromosome or entire subregion of interest, so

28 that a disease-causing polymorphism located anywhere on the chromosome or anywhere in the

29 subregion of interest will be detected by the test.

30

31 Step 2

32 It is well known that increased disequilibrium between a marker and linked disease

33 locus increases evidence for linkage provided by association-based linkage tests such as the

34 TDT. However, what has not been recognized is that the specific allele frequencies of the

35 marker locus can also have an enormous impact on the strength of evidence for linkage. I

1   showed this by analyzing equation 2 for P_t. Specifically, when a bi-allelic marker is in linkage

2   disequilibrium with a bi-allelic disease locus, the strength of evidence for linkage provided by

3   the TDT is *greatly* increased if the bi-allelic marker and bi-allelic disease locus have similar

4   allele frequencies.

5   This phenomenon is illustrated by Table 2 and explained above. For example, suppose

6   as noted above, that the susceptibility allele ("allele D") of a bi-allelic disease locus has a

7   frequency of 0.1 and further suppose that the disease locus is in half-maximal positive

8   disequilibrium with a bi-allelic marker ($\delta=\frac{1}{2}\delta_{max}$). As noted above, χ^2_{tdt} will equal only 3.61

9   ($p< 0.1$) if the frequency of the less common marker allele is 0.5; but if the frequency of the

10   less common marker allele is 0.2 (and hence much closer to the frequency of allele D) then

11   χ^2_{tdt} will equal 11.54, thus providing much stronger evidence for linkage ($p<0.005$).

12   *Therefore*, in searching for association-based linkage to a bi-allelic disease locus within

13   each of the aforementioned chromosomal segments (see step 1), it is crucial to identify and test

14   (e.g., by TDT) bi-allelic markers within each segment that have a broad range of allele

15   frequencies. An unidentified bi-allelic disease locus could have allele frequencies close to

16   0.5/0.5, 0.4/0.6, 0.3/0.7, 0.2/0.8, 0.1/0.9 or below 0.1/above 0.9; hence, it is crucial to test bi-

17   allelic markers with frequencies near 0.5/0.5 and near 0.1/0.9 as well as test others with allele

18   frequencies that fall at regular increments between the extremes of 0.5/0.5 and 0.1/0.9. By

19   testing bi-allelic markers with a broad range of allele frequencies that are spaced at regular

20   intervals between 0.5/0.5 and 0.1/0.9, one is assured of testing some bi-allelic markers whose

21   two allele frequencies are reasonably close to the allele frequencies of an unknown bi-allelic

22   disease locus.

23   Thus, for step 2, within each chromosomal segment, subsets of bi-allelic markers

24   should be identified. Each subset contains only bi-allelic markers having approximately the

25   same allele frequencies. For example, subset A contains only markers whose less common

26   allele has a population frequency of about 0.1. Similarly, subsets B, C, D, and E contain only

27   bi-allelic markers whose less common allele has a frequency of approximately 0.2, 0.3, 0.4,

28   and 0.5, respectively. In other versions of the invention the number of subsets is greater or less

29   than five, and the approximate allele frequency of the less common bi-allele of subsets is other

30   than about 0.1, 0.2, 0.3, 0.4 or 0.5 and is expected to be more than one decimal long since

31   allele frequencies from real populations are rarely round numbers. However, the crucial point

32   is that each subset should contain only bi-allelic markers belonging to one chromosomal

33   segment and the frequency of the less common allele of each subset member should be

1 approximately the same (i.e., the *difference* between the frequencies of the less common allele

2 of any two subset members should not exceed 0.15). Also crucial, as I emphasized above, is

3 that the *group* of subsets for each chromosomal segment represent frequencies near the

4 extremes of 0.5/0.5 and 0.1/0.9 as well as represent bi-allele frequencies between these two

5 extremes that are approximately evenly spaced as *illustrated* by the group of subsets referred to

6 above as A, B, C, D and E.

7

8 Step 3

9   In step 2, I described the importance of testing subsets of bi-allelic markers having

10 approximately the same frequencies for their two alleles. Here I further delineate the

11 characteristics of the markers that should be chosen for each subset by noting why it is

12 important that each subset contain more than one bi-allelic marker. Even though a particular

13 bi-allelic marker has allele frequencies that are similar to those of a closely linked bi-allelic

14 disease locus, the marker may not be in strong positive disequilibrium with the disease locus.

15 If disequilibrium is minimal, the marker will not show strong evidence for linkage under the

16 TDT or any other association-based linkage test, *even though the bi-allelic marker and disease*

17 *locus have similar allele frequencies.*

18   Hence, it is important that each subset contain multiple bi-allelic markers so that there

19 is increased likelihood that at least one of the markers will be in reasonably strong

20 disequilibrium with a closely linked bi-allelic disease locus. Beyond the cardinal criterion that

21 all bi-allelic markers in a subset have similar allele frequencies, an additional criteria for

22 selecting markers to belong to a subset is that the chosen bi-allelic markers *should not be in*

23 *extreme positive disequilibrium with each other.*

24   The reason for this is as follows: According to standard usage, the disequilibrium

25 coefficient (δ) is defined by the equation $\delta = f(AB) - f(A)f(B)$ where $f(A)$ and $f(B)$ may be

26 defined as the frequencies of the less common allele (denoted A and B) of two bi-allelic loci

27 belonging to the same subset and $f(AB)$ is the population frequency of the AB haplotype.

28 Since the two markers belong to the same subset, we may assume that $f(A) = f(B) = q$; hence the

29 maximum positive value of δ (denoted δ_{max}) is $\delta = q - q^2$. This maximum positive δ value (i.e.

30 maximum "positive disequilibrium") occurs if every chromosome that carries allele A also

31 carries allele B, and if every chromosome that carries allele not-A also carries allele not-B.

32 Hence, when two bi-allelic markers with similar allele frequencies are in extreme postive

33 disequilibrium with each other (i.e., δ is approximately equal to δ_{max}), the two loci provide

1 the nearly identical information with respect to their linkage and association with a third

2 polymorphism such as a disease locus. Hence one of the two bi-allelic markers would provide

3 no additional information and its inclusion in the subset would not increase the likelihood of

4 detecting linkage and association to a nearby disease locus.

5 Therefore, bi-allelic markers belonging to the same chromosomal segment and subset

6 should not only have similar allele frequencies, the δ value between *each pair* of bi-allelic

7 markers in the same subset should be substantially less than $\delta_{max} = q - q^2$. This assures that

8 every bi-allelic polymorphism belonging to the subset provides much new (i.e. non-redundant)

9 information about linkage and association to any nearby bi-allelic disease locus; thus testing

10 each bi-allelic marker in the subset would increase the likelihood of detecting linkage to a

11 disease locus.

12

13 Step4: Test for linkage

14 To test for (association-based) linkage to a bi-allelic disease locus, each bi-allelic

15 marker in each subset from each chromosomal segment is tested *individually* by using the

16 TDT, AFBAC method or other family-based linkage test. To conduct these tests for a

17 particular marker, members of nuclear families (most especially parents, and any children who

18 manifest disease) are genotyped at the marker being tested and the genotypes are then

19 evaluated according to the TDT, AFBAC method or other family-based linkage/association test

20 (for description of TDT and AFBAC, see Spielman et al, Am J of Human Genetics 52:506-516

21 (1993) and Thomson, Am J Human Genetics 57:487-498 (1995)). Alternatively, linkage and

22 association is tested for each marker in each subset from each segment by genotyping

23 individuals with disease and related or unrelated normal controls at each marker to be tested.

24 (End of set/subset example)

25 Further Information

26 (Step 3 is not essential for the operation or utility of this version of the invention. In this

27 set/subset example, the least common allele frequency subrange 0.1 to 0.5 is used. In versions

28 of the invention similar to the set/subset example, versions of the invention are operable and

29 have utility for any subrange of the least common allele frequency range 0 to 0.5. In addition,

30 rather than genotyping DNA from single individuals in step 4, in some versions of the

31 invention each marker in each subset from each segment is tested for association with disease

32 by evaluating DNA from pooled samples.)

33

1 PART 2 - Physical implementation of the above test

2

3 <u>Silicon chips or glass slides with arrays of oligonucleotides for testing bi-allelic markers</u>

4 Companies like Affymetrix[8] are using silicon chips or glass slides to genotype DNA

5 from one individual at thousands of bi-allelic markers. Each silicon chip or glass slide is

6 divided into a grid or 2-dimensional matrix that contains thousands of cells. To the surface of

7 each cell is attached multiple copies of a unique oligonucleotide whose sequence is

8 complementary (type (1))to one of the two alleles of a particular bi-allelic marker. Thus, DNA

9 from an individual who carries the allele hybridizes to the cell with substantially greater

10 affinity than does the alternate bi-allele. The degree of hybridization generates a signal and

11 enables the genotype of the individual to be inferred for that particular bi-allelic polymorphism

12 [i.e., the individual is homozygous (++), heterozygous (+-), or homozygous (--)]. In some

13 applications, it is crucial to attach oligonucleotides corresponding to each allele of a bi-allelic

14 polymorphism in adjacent cells so that the relative (i.e. local) intensity of hybridization in the

15 adjacent cells can be compared, thus facilitating inference of the individual's correct genotype

16 (++, +-, or --).

17 In using this silicon chip or glass slide technology to test for linkage and association,

18 the ideas detailed in PART 1 indicate how the oligonucleotides that are attached to the cells of

19 the silicon chip or glass slide should be chosen. To scan a particular chromosome or

20 chromosomal region for a bi-allelic disease locus, the chromosome or chromosomal region

21 should be subdivided into segments as described in Step 1 above. For each segment, subsets of

22 bi-allelic markers having the properties detailed in PART 1 above should be identified. The

23 DNA of select individuals (see "Test for linkage" - above) should then be assayed at each bi-

24 allelic marker in every chromosomal segment and in every subset of markers belonging to the

25 segment. This would be accomplished by attaching an oligonucleotide corresponding to one of

26 the marker's two alleles to a particular (i.e. known) cell on the silicon chip or slide. To

27 enhance assignment of accurate genotypes, it may also be advisable to attach an

28 oligonucleotide corresponding to the second allele of the bi-allelic marker in an adjacent cell as

29 mentioned in the previous paragraph.

30
31 **Industrial Applicability**

1 Scope of the Invention

2 While the description contains many specificities, these should not be construed as limitations on the

3 scope of the invention, but rather as exemplifications of versions of the invention. Accordingly the scope

4 of the invention should be determined not by the specific versions described alone, but also by the

5 claims and their legal equivalents and also by any future claims drawn to the invention and future

6 descriptions of versions of the invention.

7

8 Notes:

9 The present patent application is a continuation-in-part of U.S. Patent Application 09/947,768 (filed 5

10 SEPT 2001). And 09/947,768 claims priority from US Provisional 60/230570 (filed 9/5/2000). Patent

11 Application 09/947,768 is a continuation-in-part of U.S. Patent Application 09/623,068 (filed 26 AUG

12 2000). The present patent application is also a continuation-in part of Patent Application 09/623,068

13 (filed 26 AUG 2000). Application 09/623,068 claims priority from PCT/US99/04376 filed (2/26/99).

14 PCT/US99/04376 claims priority from US Provisional applications: 60/076182 filed 27Feb1998,

15 60/086947 filed 27May1998, 60/076102 filed 26 Feb 1998 and 60107673 filed 7 Nov 1998. Each of the

16 following patent applications are incorporated herein by reference in their individual entireties: U.S.

17 Provisional Patent Application 60/230570, PCT/US99/04376, U.S. Patent Application 09/623,068, and

18 U.S. Patent Application 09/947,768.

19 Each of the following patent applications are also incorporated herein by reference: US Provisional

20 applications: 60/076182, 60/086947, 60/076102 and 60107673.

21 The reader's attention is directed to the following papers each of which is open to the public and each

22 of which is incorporated by reference herein in their entirety: (1) McGinnis, Ewens & Spielman, Genetic

23 Epidemiology 1995 ; 12(6) : 637-40. (2) RE McGinnis Annals of Human Genetics vol 62, pp. 159-179.

24 The papers in the endnotes below are incorporated herein by reference (3) Unpublished manuscript

25 Detection of linkage: Comparison of the affected sib pair (ASP) test and transmission/disequilibrium test

26 (TDT) included with each of one or more US Provisional applications: 60/076182, 60/086947,

27 60/076102 and 60107673.

28

---

[1] Weighing DNA for Fast Genetic Diagnosis, Science, March 27, 1998, vol. 279, pp. 2044-2045.

[2] Spielman, R.S. and Ewens, W.J. The TDT and Other Family-Based Tests for Linkage Disequilibrium and Association , American Journal of Human Genetics, 59: 983-989, 1996.

[3] "Mathematical Theory of Optimization" The New Encyclopedia Britannica, 15th edition, vol. 25, pp. 217-221.

[4] American Journal of Human Genetics, vol. 57: 439-454, 1995.

[5] American Journal of Human Genetics, vol. 58: 1347-1363, 1996.

[6] Human Heredity, vol. 44, pp. 225-237, 1994.

[7] Human Heredity, vol. 46, pp. 226-235, 1996.

[8]: Accessing Genetic Information with High-Density DNA Arrays, Mark Chee, et al. Science, vol 274, Oct. 25, 1996 , pp. 610 – 614.

[9] Large Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome, Wang, et. al., Science, May 15, 1998, vol 280, pp. 1077-1081.

[10] (1)Schuster, H. et al (1995) Nature Genetics, 13(1) : 98 – 100.
(2)Gyapay, G. et al (1994) Nature Genetics, 7: 246-339.
[11] Some versions of oligonucleotide technology and it's uses to obtain genetic information are included in the following papers:
(1) Accessing Genetic Information with High-Density DNA Arrays, Mark Chee, et al. Science, vol 274, Oct. 25, 1996 , pp. 610 – 614.
(2)Genetic analysis of amplified DNA with immobilized sequence- specific oligonucleotide probes, Saiki,et al. Proc Natl Acad Sci USA vol 86, pp.6230-6234.
(3)Allele-specific enzymatic amplification of β-globin genomic DNA for diagnosis of sickle cell anemia, Wu, et al., Proc Natl Acad Sci USA vol 86 pp 2757-2760.
(4)Automated DNA diagnostics using an Elisa-based oligonucleotide ligation assay, Nickerson, et al., Proc Natl Acad Sci USA vol 87, pp. 8923-8927.
(5)Genetic analysis of amplified DNA with immobilized sequence specific oligonucleotide probes, Saiki, et al., Proc Natl Acad Sci USA vol 86 pp 6230 – 6234.
[12] Padlock Probes:Circularizing Oligonucleotides for Localized DNA Detection, Science, Sept. 30, 1994, vol. 265, pp. 2085-2088.
[13] SNP attack on complex traits, Nature Genetics, Nov. 1998, vol. 20 no. 3, pp. 217-218.